Application of Bayesian Networks to Problems within Obesity Epidemiology

A thesis submitted to the University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Medical and Human Sciences.

2011

Nicholas J Harding

School of Medicine

Faculty of Medical and Human Sciences

Contents

Word Count: 44,860

List of Figures

List of Tables

Abstract

Obesity is a significant public health problem in the United Kingdom and many other parts of the world, including some low-income settings. Although obesity prevalence has been rising for several decades, governments have been slow to implement policies that may have an impact at a population level. Numerous socio-demographic factors have been linked with obesity, but are highly intercorrelated, and identifying relevant factors or at-risk population groups is difficult.

This thesis uses a graphical modelling approach, specifically Bayesian networks, to model the joint distribution of socio-demographic factors and obesity related behaviour. The key advantages of graphical models in this context are their ability to model highly correlated data, and to represent complex relationships efficiently as network structure.

Three separate pieces of work comprise this thesis. The first uses a sampling technique to identify the networks that best explain the observed data, and employs the common structural features of these networks to infer conditional dependencies present between socio-demographic variables and obesity related behaviour indicators. We find determinants of recreational physical activity differ between males and females, and age and ethnicity have a significant influence on snacking behaviour. The second piece of work uses Bayesian networks to build a model of health behaviour given socio-demographic input, and then applies this to data from the 2001 census in order to provide an estimate of the health behaviour of a real population. The final analysis uses Bayesian network structure to explore potential determinants of body fat deposition patterns and compares the results to those derived from a Generalized Linear Model (GLM). Our approach successfully identifies the main determinants, age and Body Mass Index, although is not a genuine alternative due to a lack of sensitivity to less important determinants.

Beyond the application to obesity, results of this thesis are of a wider relevance to epidemiology as the field moves towards an increased use of Machine Learning techniques. The work conducted has also met and overcome several technical issues that are likely to be of relevance to others exploring similar approaches.

University of Manchester

Application of Bayesian Networks to Problems within Obesity Epidemiology

A thesis submitted for the degree of Doctor of Philosophy

Nicholas J Harding

8th April, 2011

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Acknowledgements

Completion of this thesis represents not just my own effort, but that of several others also. I would like to thank David Hoyle for countless hours of patient explanation and problem solving. Iain Buchan for providing the opportunity and direction. Staff within NIBHI particularly Paul Jarvis, Matt Sperrin and Nathan Green for programming, statistical and LaTeX help and advice. Finally, Anna, my parents and my brother for continued support, belief and encouragement. Thanks all.

Glossary

Arc/Edge: Directed connection between two nodes representing a conditional dependency.

Beta Distribution: A probability distribution representing the probabilities of the outcomes of a discrete binomial event. It is a special case of the Dirichlet distribution where the random variable is binomial.

Child node: The node to which a directed arc is incident from a parent node.

Dirichlet Distribution: A probability distribution representing the probabilities of each of several possible outcomes of a discrete event. It is the multinomial extension of the Beta distribution.

Edge relation feature: A probability estimate of an arc existing between two nodes following approximation over the space of of all possible network topologies.

Likelihood: The retrospective probability of the observed data.

Markov Chain-Monte Carlo: An approach often used to sample from probability distributions. A Markov Chain is a memoryless random process, *i.e.* its future states depend only on its current state. Monte Carlo algorithms use repeated samples to approximate a distribution.

Metropolis-Hastings sampling: A Markov Chain-Monte Carlo technique for approximating distributions that are difficult to evaluate analytically.

Move: An operation used to generate a new network topology from an existing network topology.

Node: A vertex representing a random variable in a Bayesian network.

Parameters: The probabilities associated with the outcomes of a child node associated with each possible combination of parent node outcomes.

Parent node: The node from which a directed arc is incident to a child node.

Parentset: A collective term for the parents of a given node.

Posterior: The distribution of a random variable in Bayesian statistics, composed of the Prior distribution and the likelihood function of the data.

Prior: The distribution assigned to a random variable in Bayesian statistics before the incorporation of data.

Probability valley: A region of low probability space, which consequently is unlikely to be visited by a Metropolis Hastings algorithm or a similar approach.

Pseudocount/Hyperparameter: These are the parameters of the Prior distribution, they influence the posterior in an identical manner to counts, but are not actually observed.

Topology: The structure of a network in terms of nodes and arcs.

Acronyms and Abbreviations

BN: Bayesian Network.

BMA: Bayesian Model Averaging.

BMI: Body Mass Index.

DAG: Directed Acyclic Graph.

ERF: Edge Relation Feature.

HSE: Health Surveys for England.

IPA: Incidental Physical Activity.

MCMC: Markov Chain-Monte Carlo.

MR move: Multiple Reversal move.

NS-SEC: National Statistics Socio-Economic Classification.

OPA: Occupational Physical Activity.

PDAG: Partially Directed Acyclic Graph.

REV move: Grzegorczyk-Husmeier edge reversal move.

RPA: Recreational Physical Activity.

WHR: Waist to Hip Ratio.

Chapter 1

Introduction

This chapter introduces the themes and concepts explored in the thesis. First the severity of the obesity epidemic in the United Kingdom and elsewhere is established, and methodological problems that epidemiologists face are outlined. Obesity in the UK has been rising consistently for several decades, despite this, very little is known regarding the potential effectiveness of population level interventions. One of the primary reasons for this is the lack of understanding of the root causes of behaviour that results in obesity. A pertinent problem is the fact that many factors related to obesity are highly correlated, making identification of relevant factors difficult, this thesis applies statistical modelling techniques to help overcome this issue. This chapter provides an brief introduction to these techniques and other related approaches in epidemiology. The final part of the current chapter provides a detailed overview of the structure of the thesis.

1.1 The Obesity Epidemic

Obesity is the condition of having excess body fat, which is associated with increased health risks. The Body Mass Index (BMI, kg/m^2) is an imprecise but useful measure of body fatness, or adiposity, and is typically used to define obesity (BMI$\geq$ 30) and overweight ($25 \leq$ BMI ≤ 30) [1].

Obesity is essentially the result of energy imbalance between that consumed and that expended, with excess energy being stored as fat. The link between energy imbalance and obesity is complex, with significant genetic and metabolic components. Obesity prevalence has risen sharply in the UK in recent years, attracting the commonly applied label ‘epidemic’; figure 1.1 shows 23.6% of males and 24.4% of females were obese in 2007, up from 13.2% and 16.4% in 1993 [2].

This epidemic is the result of a complex environment that encourages people to consume more and exercise less [3]. Despite the insistence of the food industry to the contrary [4], the available evidence suggests people are consuming more than ever [5]. This increased consumption is due to a multitude of social and economic factors, culminating in an environment with high availability of cheap, energy dense, highly palatable food that is aggressively marketed to the population. In addition, the current environment provides less opportunity for physical activity, with shifts in the labour market and technological advancement promoting a more sedentary lifestyle [5]. A myriad of factors interact to create the obesogenic environment, the relative importance of which varies between individuals and groups. Although the magnitude of the obesity epidemic has long been recognised, finding

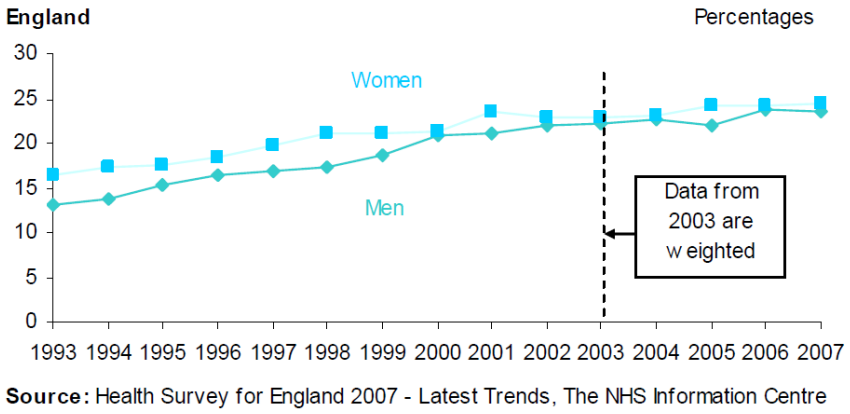


Figure 1.1: Prevalence of obesity in UK adults by gender, 1993 to 2007

a solution is likely to be one of the biggest public health challenges of the 21st Century [6]. An approach that is able to model interacting factors represents an important step to the understanding of obesity.

The public health and economic implications of rising adiposity across society have led to widespread concern [7, 8]. Obesity prevalence has been rising year on year for several decades. Obese individuals have a significantly reduced life expectancy [9, 10], and are at a higher risk of diabetes, cardiovascular disease [11], cancer [12], musculoskeletal conditions [13], and numerous other health problems [14]. The annual direct cost to the National Health Service (NHS) of obesity and its consequences has been conservatively estimated at approximately £1bn in 2001 [15], around 2% of total expenditure, and will rise with increasing obesity prevalence. This figure does not include cost attributable to overweight, which has been tentatively estimated as equal to the cost of obesity [15]. These numbers, however, belie the true cost, with the indirect costs due to absenteeism and reduced productivity estimated at a substantial proportion of the direct cost.

Childhood obesity is a major public health problem, with obesity prevalence rising from 11.1% and 12.2% to 16.8% and 15.2% in boys and girls respectively in the UK from 1995 to 2008 [2]. Obese children are more likely to become obese adults [16, 17], with associated health problems and subsequent costs. Long term health implications of obesity from childhood are unclear, but may be even more severe as excess body fat will have been carried for a longer period than most people currently obese.

Risk of developing type-II diabetes is closely associated with rising BMI [18–20]. The prevalence of diabetes is increasing, with obesity responsible for a large proportion of cases [21]. Recent years have also seen a surge in diagnoses in children and adolescents [22]. Obesity associated type-II diabetes is becoming increasingly common in children and adolescents [23], and some may experience secondary complications [24]; diabetes may afflict a third of US children if current trends continue [22]. Diabetes and its subsequent complications represent a significant financial cost in terms of drug treatment and complications of disease [25,26]. Increasing rates of obesity coupled with the UK's aging population [27] will allow prevalence of diabetes to continue to rise at huge financial and human cost [28,29]. Cardiovascular disease is the leading cause of death in the UK [30], and obese individuals are vulnerable to a much higher risk [11]. Although death rates from coronary heart disease (CHD) and all circulatory diseases have been declining for three decades [30], obesity is becoming increasingly important as a risk factor [31]. Reductions in overall death rates are attributed to improving treatments and management of other risk factors, such as smoking, hypertension, and cholesterol levels [31]. A modelling study attributed $\sim 7,700$ CHD deaths (of $\sim 100,000$) in 2001 to increases in obesity, diabetes and physical inactivity since 1981 [31]. Obesity is associated with increased risk of several cancers, including endometrial, kidney, gall bladder (in women), breast (in post menopausal women), liver and colon (particularly in men) [32–36]. Links between obesity and cancers of the prostate, pancreas, ovaries and haematopoietic cancers are less well defined [37–41]. Obesity contributes to a substantial proportion of cancer mortality; Renehan and colleagues [12] estimated the fraction attributable in the United States at 1.5-3.6% in men and 2.3-5.9% in women. Although obesity is currently behind smoking and alcohol abuse as the main cause of cancer mortality in developed countries [42], it still represents a major economic and human burden which is likely to rise given further increases in obesity prevalence.

Treatment of obesity and its consequences represent direct costs. Indirect costs also attributable to obesity include lost earnings due to premature mortality and certified sickness, absenteeism, lost productivity in the workplace, and incapacity allowances. The 2004 House of Commons Health Committee Report on Obesity [15] estimated the indirect cost of obesity at £2.35-2.6bn in 2002; this is derived from £1.05-1.15bn from lost earnings due to premature mortality, and 1.3-1.45 bn from certified sickness. This figure represents approximately two-thirds of the total cost attributed to obesity in the UK, and may be an underestimate due to exclusion

of absenteeism costs, reduced productivity and incapacity benefit (though the latter is likely to be small, approx £9m in 2005/6 [43]). Other studies have estimated the indirect cost of obesity at $23.0bn of $68.8bn ($\sim$ 33%) in 1990 [44], $47.6bn of $99.2bn ($\sim$ 48%) in 1994 [45] and $56bn of $117bn (48%) in 2000 [46] in the United States, and $637m of $\$1,721$m ($\sim$ 37%) in 2005 in Australia [47]. Although the proportion of the UK total cost is larger, this is likely to be due to different methods and inaccuracies in estimating direct costs. These indirect costs reflect not only increased mortality of obese individuals, but a higher tendency to take absence, both certified and uncertified [48–50], and lower productivity in the workplace [51]. Not all of the negative implications of rising obesity manifest fully in financial costs; obesity is a major contributory factor to depression in some individuals [52], and is generally associated with a reduced quality of life [53].

In order to produce estimates of future costs of obesity the government commissioned the Obesity Foresight[1] report, which undertook a simulation that modelled future costs of rising obesity prevalence. The model was based on an extrapolation of the observed rise in obesity in the UK from 1993-2004. A microsimulation was carried out that simulated mortality and disease in the population. Risk of disease was based on a literature review where the odds ratio of a number of obesity related disease was estimated. Cost was linearly projected from 2004 NHS cost data, *i.e.* if the number of diabetes patients doubled, the cost to the economy would also double. A figure of almost £50bn by 2050 was reached, however this work is based on a crude linear extrapolation, and is not widely accepted.

1.2 Challenges in Obesity Epidemiology

The cause of obesity at an individual level is simple; a mismatch between personal energy intake and energy expenditure leading to the storage of excess energy as fat. However, understanding the factors that affect or determine energy intake and expenditure at a population level is far from straightforward. The link between social demographic factors and obesity is well studied and frequently discussed, but there is no simple association [54–56].

Obesity is often depicted as affliction of the urban poor, but the reality is far more nuanced. Data has shown that the median as well as the mean BMI is increasing, suggesting weight gain across the entire population [2]. Improved characterisation of the obesogenic environment will allow the identification of factors

[1] www.foresight.org.uk

that determine and mediate obesity in populations. Such information will allow the design of better targeted and efficient interventions, and the identification of population groups at particular risk. The obesogenic environment is highly complex, and is likely subject to many interactions, whereby factors may have a strong influence in some population groups but little in others. The identification of 'pressure points' where interventions can be shown to be cost-effective is needed, this requires a much more complete understanding of the environment.

Obesity is strongly tied to socio-demographic factors, albeit in a complex manner. Several papers have noted variation in diet quality with socio-economic status, income and education [57–62]; poorer areas have also been shown to have a higher density of fast food outlets [63, 64]. Numerous studies have reported that participation in recreational physical activity is associated with socio-demographic factors [65–69]. Access to amenities such as green spaces and leisure facilities may vary with deprivation [70–73], although it is difficult to establish a causal link between access and behaviour. Risk factors for energy imbalance are not uniformly distributed and are likely be clustered in population groups [74].

Although each of these factors exhibit links to obesity, such socio-demographic variables are highly and inextricably correlated. Much of epidemiology uses regression analyses to explain the variation in data, and to attribute it to various independent exposures. In this context, however, the concept of an independent risk factor for obesity is nonsensical; socio-demographic factors are highly correlated, which can be a cause of bias in traditional regression models, which may not adequately adjust for highly correlated covariates [75]. In regression analyses, correlation between covariates can provide misleading results; a genuine effect may be obscured by intracorrelation between several variables as correlation between covariates is not reported by regression models. This thesis argues that in such cases, epidemiology should move beyond modelling exposure-disease relationships, and attempt to model entire systems, including dependencies between covariates.

Population data is necessary to explore the relationships between factors associated with levels of energy intake and expenditure. However, measuring levels of energy intake and expenditure has proved difficult beyond controlled laboratory conditions. Most epidemiological studies are observational in nature, as this is the only practical method of obtaining data at a population level. Such data is subject to bias from a wide range of sources, such as uncontrolled confounders, selection and participation bias, measurement error and missing data. To date, there is very little information available on the health behaviour of populations. Although obesity re-

lated behaviours are known to vary with numerous socio-demographic factors, data is not available regarding energy consumption and energy expenditure levels at a population level. Consequently, the scale of overconsumption and sedentariness are unknown. The discrepancy between energy consumed and energy required is termed the *'energy gap'*. It is the widening energy gap that is responsible the observed rise in obesity, and is therefore the obvious target for intervention. However, the size of the energy gap is unknown. A number of papers and editorials refer to an energy gap of 100kCal/day being responsible for the obesity epidemic [76–78]. Morabia and Costanza [78] speculate that the obesity epidemic could be eliminated if all individuals in a population walked an additional 15 (brisk pace) to 60 (slow) minutes a day to burn the additional 100kCal required. This figure is Hill's 2003 estimate [79] of an energy imbalance sufficient to explain the observed rise in body weight in 20-40 year old Americans over 8 years. However, as Swinburn et al. [80] point out, this estimate does not take into account the extra energy cost of maintaining the additional body tissue, revealing the figure to be a substantial underestimate. Hill's calculation is based on a 1958 paper by Wishnofsky [81], which states that an energy imbalance of 3500kCal ($\sim 32.2 MJkg^{-1}$) is required to lose or gain 1 pound (0.455 kg) of body weight. Although a reasonable approximation [82], the reality appears significantly more complex, with numerous other factors affecting the energy deficit required for weight loss [82, 83]. Several papers have attempted to uncover the relationship between energy intake, expenditure and weight gain, some based on the energy content of fat mass [84, 85], others on data from observational studies [80]. However, there has been relatively little work to estimate the degree of energy imbalance in current populations. It is known that to counter the obesity epidemic we must close the energy gap, but it is not known how large this gap is. This is obviously a major barrier to intervention design; numerous small interventions may be a viable solution, or futile.

BMI is currently accepted as the primary measure of obesity within populations. Its main advantage lies in its relative ease of measurement when compared to alternatives such as skin-fold thickness or percentage body fat. BMI has been shown to be an unreliable measure of individual adiposity [86], which is especially true for tall individuals, where additional height is not fully accounted for. People of a certain BMI are at varying risks due to adiposity, which the BMI measurement fails to capture. The Waist Hip Ratio (WHR) is a proxy measure of the visceral fat to body size, which can predict risk of cardiovascular disease independently of BMI [11]. Nonetheless, BMI is a satisfactory measure of excess weight in popula-

tions rather than individuals.

Adiposity indicators are insufficient for identifying relationships between social factors and obesity. The relationship between energy intake, energy expenditure and weight gain is highly complex. As a result, inferring behaviour from weight status or vice versa is not possible. Throughout this thesis, I investigate the influence of social factors on behaviour, rather than weight status. Given behaviour, weight status is independent of socio demographic factors. In addition, individuals' behaviour may not be in equilibrium with their weight. If health behaviour is reasonably constant over time, body weight will naturally move towards an equilibrium state between energy consumed and expended- energy expenditure increases with body weight (though the reality is complex). Individuals exhibiting unhealthy behaviours will not immediately exhibit the full results of that behaviour. It is not known how many individuals are currently in an equilibrium state, and therefore how reflective BMI or weight indicators are of current behaviours and future body weight. Health behaviours are the natural targets for intervention, as they are the fundamental cause of population obesity. Several behavioural indicators can be examined simultaneously, allowing better characterisation of the obesogenic environment, and opening up the possibility of revealing direct relationships between behavioural indicators.

1.3 Obesity Interventions

A large number of studies have investigated the effects of various behavioural weight loss strategies in adults [87–90]. These strategies typically focus on changing individual habits that predispose to overweight and obesity. Common components of interventions include providing nutritional education, improving diet, and increasing levels of physical activity. The majority of studies are reported to be effective [90], but this may be due in part to publication bias. An insufficient period of follow up is a limitation of many intervention studies [90] with few exceptions [91, 92]. Even in longer term studies usefulness is limited by high attrition and loss to follow up. No long term studies have reported anything more than a modest weight reduction [88, 90].

Comprehensive reviews indicate that weight loss from behavioural interventions tends to peak around 6 months, followed by a period of regain, though not all studies have sufficient follow up periods to show this [88]. Such weight rebound is presumably due to at least partial reversion to previous habits. Maintenance of

weight loss is difficult [93], maintaining a lower weight requires a lower energy intake than original weight and any deviation from habits imposed by intervention will create an energy gap and subsequent weight gain. People may be unwilling or unable to make sufficient long term lifestyle changes to maintain weight loss, even if initially enthusiastic towards the intervention. Intervention efforts focus much more on weight loss rather than maintenance of a healthy weight [79]. Only a few studies have investigated maintenance of weight [94]; Hill and colleagues argue that efforts to maintain weight should be prioritised [79].

Intervention studies require active participants, and participation indicates a willingness to change that predisposes to success. Few studies have published data on uptake rates, and such data is not collated in reviews [87–90]. As withdrawal constitutes a failure, levels of attrition are crucially important in understanding the effectiveness of an intervention; but are not given sufficient weight in reports. The majority of studies focus on average weight loss and proportions of participants achieving a percentage loss of original bodyweight. However these figures make no adjustment for those who decline further treatment; this may result in a skew towards intensive interventions when gauging effectiveness. Behavioural interventions show great diversity in setting, participants, intensities, methodologies and scales [88, 90]. Forces of obesity acting differ according to social and cultural contexts, and different interventions will be suited to different situations. This emphasises the importance of characterising the specific obesogenic environment of groups when constructing interventions.

Recent opinion suggest solutions for the obesity epidemic lie in environmental changes rather than individual behavioural change [95–98]. Although obesity is fundamentally derived from the simple formula of energy imbalance, the environment that creates it is complex. Despite apparent consensus on the ineffectiveness of individual based interventions and the need for policy change, there is limited evidence regarding population level interventions. Some studies have carried out interventions at the community level but with little success [99, 100] possibly due to ineffectiveness of solitary policy changes. Numerous sources have suggested possible policy interventions as a matter of urgency, including among others; taxes and advertising restrictions on unhealthy food [101], improved access to nutritional information [102] and changes in the built environment [103]. Although such policies are sensible and may play a significant role in future strategy, solid evidence of effectiveness is required before substantial investment can be made. There is very little evidence to suggest if or how these changes would influence population

eating and physical activity habits. Evidence of the ability of these interventions to influence national obesity prevalence and cost effectiveness is even further away. Although the cost of a potential intervention may be small in comparison to the cost of obesity, effectiveness is not assured [104]. Research effort has focussed on cost and cure; but research is needed to understand the effects of change.

Hill and colleagues [95] discuss the lack of evidence available for policy interventions and divide approaches into four categories: information, accessibility, price and marketing. A detailed review of interventions in each of these four categories is included in appendix A.1.

1.4 Introduction to Graphical Models

As discussed in the previous section, the study of obesity generates complex problems, involving the interaction of numerous correlated variables. The utility of standard regression models commonly used in epidemiology is limited, as it is the joint distribution of variables that is of interest, rather than the effect of a set of variables on a single output. A complete description of graphical models and Bayesian networks can be found in Chapter 3.

In October 2007, at the beginning of this PhD program, the Office of Science released the long awaited Foresight report into obesity, with the aim *'to produce a long-term vision of how we can deliver a sustainable response to obesity in the UK over the next 40 years'*. At the centre of this lengthy report is a large graphic that represents the linkages between factors contributing to the obesity epidemic, an obesity 'system map'. This was constructed as a collaborative effort by a number of researchers in the field, and compiled from numerous reviews. The intention of the map is to provide an insight into the range of reinforcing causal factors that contribute to individual obesity.

In a *Lancet* editorial, Jack [105], criticises the map as 'convoluted' and states 'In a very complex way, the graphic simply makes the banal point that the many factors that contribute to obesity all influence each other'. The obesity map is reproduced in figure 1.2[2], and it is certainly complex. However, the fact that many of the relevant factors interact is far from banal- it represents a major challenge to epidemiology. A visualisation such as this allows us to see the entire picture, a picture that could not be given by regressing any number of factors on some variable

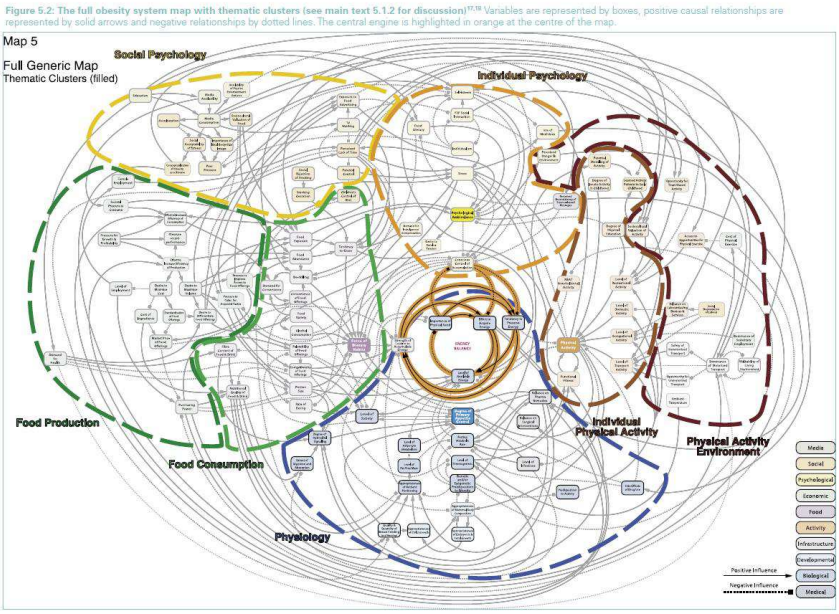
[2]Image has been reduced, a high resolution interactive version is available from `http://www.shiftn.com/obesity/Full-Map.html`

representing obesity. Graphical models provide an opportunity to incorporate and visualise this complexity.

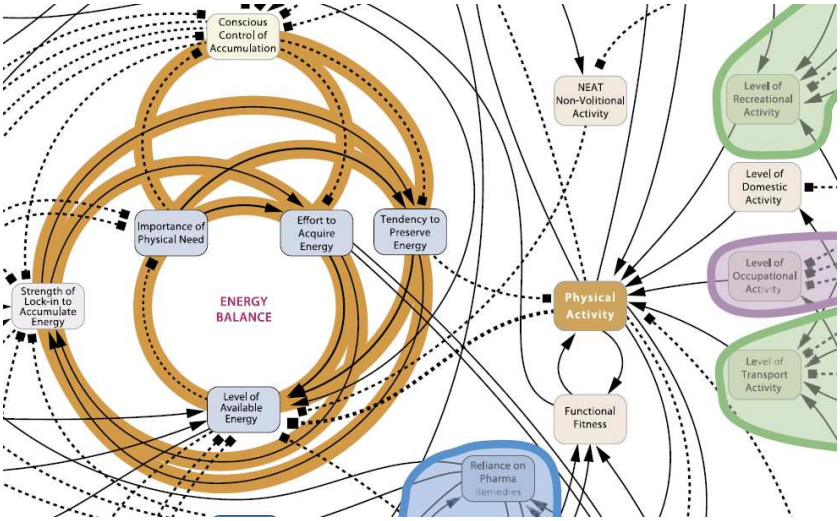
Graphical models are a representation of a probability distribution in graphical form [106]. They have many applications, and have been used in a wide variety of fields, such as gene expression [107] and decision systems [108]. Graphical models have several features that make them an attractive modelling tool in this context. Models of entire systems can be built, rather than modelling the outcome of a single variable. Numerous variables of interest and their interdependencies can be examined simultaneously, which is desirable for complex models of obesity dynamics. Bayesian networks are a type of graphical model, where the joint probability distribution is expressed in Bayesian form. A key property of Bayesian networks is the ability to score the structure of a network by calculating the likelihood of the observed data given that network. This thesis applies this property to find structure in datasets of obesity related variables.

The conditional dependencies present in a graphical model are represented by arcs between nodes representing random variables. This facilitates understanding of the statistical distribution represented by the model. This property makes graphical models a useful tool for identifying structure within data. Correlations and codependencies are clearly represented in the model output- this is advantageous if results are to be interpreted by individuals without a statistical knowledge.

Graphical models are commonly used within Machine Learning (ML); a large field involving the computer application of often probabilistic models to data. Substantial literature exists regarding learning structure of graphical models from data, which can provide information about the conditional dependencies present. In this thesis a sample of the most probable networks given the observed data is derived. The resulting networks provide information regarding structure present in the data. Although graphical models can incorporate expert knowledge, the approach taken here is entirely data driven, or *unsupervised*. Investigator bias is an issue in epidemiology, particularly where a large number of models are fitted in the exploratory stages of an analysis. Often researchers visit a study or dataset with a hypothesis in mind; the existence of a particular mindset may bias the scientific method, perhaps even unconsciously. Despite semi-formalised model selection methods the preference of an investigator may still have a substantal impact on the final model fitted. Data driven ML techniques can identify and select models without problems of investigator bias.



(a) Overview of map



(b) Detail of map

Figure 1.2: Obesity system map from the Foresight report

1.5 State of the Art

Epidemiology has not yet fully embraced Machine Learning techniques, although they are becoming common in related fields such as medicine [108, 109] and systems biology [107]. The most common applications for such techniques are still data intensive fields, such as economics [110] and genetics [111, 112]. However, the increasing digitalisation of health records and the subsequent high volumes of data mean it is likely that data driven approaches will be an important tool for epidemiologists in the coming years.

The following sections provide a brief overview of problems related to epidemiology to which machine learning approaches have been applied. Most current applications of machine learning in epidemiology involve classification, which is applied to latent variables or evaluation of risk.

Latent Class Analysis

Latent variables are variables that cannot be directly observed, but instead are inferred using statistical models. These variables may correspond to a real variable that cannot be measured for reasons of practicality, or they may represent more abstract concepts, such as mental or behavioural states.

A paper by Simpson *et al* [113] used the Microsoft library *INFER.NET*[3] to perform Bayesian Inference via a Hidden Markov Model (HMM) to identify latent classes of atopic sensitivity in children, based on Immunoglobin antibody responses at different ages from the Manchester Asthma and Allergy Study [4]. An HMM is a type of graphical model where the state of some unobserved (*i.e.* hidden) variables are inferred from a set of observed variables; in this case atopic classes from immunoglobin sensitivity measurements. The Markov property derives from the transition of children between sensitisation classes with age. Although essentially unsupervised, the investigators must declare the number of latent classes in the model. The results represent a significant departure from the existing paradigm of atopic vs non-atopic children. The 'multiple early' sensitisation class showed much poorer lung function and airway reactivity than other classes. Following characterisation of children in each class, these were described as 'multiple late onset', 'dust mite', 'non dust mite', and 'no latent vulnerability'.

Li *et al* also apply machine learning to identify latent classes among cases of

[3] `http://research.microsoft.com/en-us/um/cambridge/projects/infernet/`
[4] `www.maas.org.uk/`

Glioma using molecular data [114]. A technique known as non-negative matrix factorization was applied to RNA sequences taken from tumour tissue. Six distinct glioma subtypes were identified.

Both of these examples show the utility of machine learning techniques are capable of grouping similar observations from complex data in a manner that may not be intuitive to investigators. Results may be more clinically valid than typical presumptions such as onset age or visual appearance.

Classification and Prediction

Within a medical setting, the use of machine learning, especially the use of network based technologies, represents a novel way of combining information from a variety of sources to obtain a clinically useful result. Medical data can often be complicated by missing values and non linear relationships- machine learning technologies are well suited to these problems. Several studies have used Artificial Neural Networks (ANNs) to perform a variety of classification tasks. ANNs are another example of a graphical model, and are based upon the architecture of biological neural networks. They are made up of a large number of artificial neurons that are highly connected, *i.e.* receive input from and output to a large number of other neurons. A neural network generates an output, or set of outputs based on inputs. A response given input is learned by adjusting the sensitivity of each neuron in the model given input from adjacent neurons. As well as this 'tuning', the structure can also be altered to improve performance. The distributed structure of ANNs provides a solid framework for dealing with non linear data, and has found many applications. Hummel *et al* [115] use ANNs to predict risk of rejection in kidney transplants, while Caocci *et al* [116] predict risk of rejection of stem cell transplants in thalassaemia patients, and Llorca *et al* [117] predict risk of mortality after lung transplantation. In each case a wide variety of inputs are provided to the model, parameters and structure are then learned according to a training set. The final binary output can also be associated with a probability or certainty measure. There are other examples of the application of ANNs to predict risk of complications following major surgery, and in some cases has been empirically shown to be an improvement upon the more conventional technique of logistic regression [117, 118]. ANNs have also been used in epidemiological settings to classify occupational exposure to chemicals [119].

A less sophisticated method for classification is the use of *Classification and Regression Trees* (CARTs). A CART model uses a training set that contains obser-

vations and pre-assigned categories. The model attempts to subdivide the observed data into classes. Such models represent an efficient method of categorising data, and the technique has been applied in a wide range of disciplines- from models of patient cost [120] to medical decision making [121] and identifying children at high risk of future obesity [122]. The criteria used by CARTs to categorise data may be informative in some cases, for example what level of some factor constitutes risk. However the approach has a number of serious limitations, such as reliance on hierarchical operations and difficulty of interpretation [123]. Unlike graphical models, the structure of the CART does not represent relations between factors under investigation.

Use of Graphical Models to Identify Conditional Dependencies

As noted, the structure of a graphical model provides information regarding the conditional dependencies present in a model. Given data the structure of a Bayesian network can be scored, *i.e.* the likelihood of the data given the network can be calculated, hence the relative value of each network can be computed. By identification of common features shared by the most probable networks, dependencies within the dataset can be determined. Furthermore, a Bayesian network can distinguish between correlation and direct dependence [124]. However, as the number of possible networks is vast, techniques that approximate the distribution must be used. Applications of this technique include identification of gene expression relationships [107, 125], gene regulatory systems [126], identification of muscle synergies in physiology [127], and metabolic and neural networks [128].

The approach taken in this thesis is markedly different to these applications. The models presented here attempt to find structure in data in an unsupervised manner. The author conducted a methodical literature search, and found no papers using a data driven approach to identify interdependencies between variables in epidemiological data A.2. Such methods are well known in gene expression [107] and some other applications.

1.6 Aim and Objectives

Aim: ***To apply Bayesian networks to typical problems in obesity epidemiology and to evaluate their utility in this context.***

Specifically, this thesis applies graphical modelling techniques to three separate problems in the field of obesity. In each case, I model Health Surveys for Eng-

land data using structure of Bayesian networks to represent the interdependencies present. It is anticipated that as well as achieving a solution to each problem, I will be able to evaluate the utility of structural graphical modelling as a tool for epidemiologists working with complex common datasets. A detailed overview of these problems and a summary of the contents of the thesis are provided in the following sections.

1.6.1 Overview of Obesity Problems Tackled

This section details the three applications of Bayesian network modelling contained within this thesis. Applications were selected with considerations of both identifying a topic important in the context of obesity epidemiology, and also one that would showcase the utility of a graphical modelling approach.

Application 1: Using Bayesian Networks to Identify Factors Influencing Health Behaviour

Attempts to characterise the relationships between socio-demographic factors and obesity related behaviours are made difficult by the highly correlated nature of social factors. This study uses Metropolis Hastings sampling to obtain a number of Bayesian network models of relevant variables in Heath Survey for England 2003 and 2006 data. The presence of structural features of these graphs are used to make inferences about the relationships present within data. This approach moves away from single response models (*i.e.* regression analyses), and allows us to investigate relations between several obesity related behaviours simultaneously. By identifying such relationships, it is hoped this model will be able to generate or shape hypotheses for further investigation and to identify targets for future obesity interventions.

Application 2: Construction of a Bayes Classifier to Predict Health Behaviour in Local Populations

Health behaviour in populations is difficult to measure and estimate. National population surveys such as the HSE are limited by under-sampling of small areas or sub-populations and by participation bias. This study seeks to use HSE data to estimate levels of various obesity related behaviours in a real sub-population. A Bayesian network model of obesity related behaviour given socio-demographic factors is built from HSE data. Data on socio-demographic factors from the Greater

Manchester conurbation in the 2001 census is applied to the model, resulting in an estimate of obesity behaviours in the Greater Manchester population.

Application 3: Learning Bayesian Networks to Identify Predictors of Waist Hip Ratio in UK Adults

Waist Hip Ratio (WHR) is an important predictor of obesity related disease, independently of BMI. Factors influencing fat deposition are not well understood. Several studies have highlighted smoking as a potential contributor to fat deposition patterns, however smoking exhibits close correlation with a number of other potential risk factors, such as physical activity and alcohol intake, which may confound this link. In this study I use the structure of graphical models to identify potential determinants of WHR. Results are compared to those derived from a generalized linear model.

1.6.2 Thesis Structure

This introductory chapter (1) aims to provide a compelling argument of the severity of the obesity epidemic, and the need for better tools to understand it. In addition, it discusses issues currently problematic for policymakers, such as the poor characterisation of the obesogenic environment, the lack of information regarding population health behaviours and the difficulties associated with a single outcome measure for obesity. Details of the data used in this thesis are provided in Chapter 2. The brief introduction to graphical models in the current chapter is expanded in Chapter 3. This chapter discusses Bayesian networks in detail, and outlines how the probability of a particular network can be evaluated given data. The chapter also describes how underlying Bayesian theory can be applied so networks can communicate information about the interdependencies present in the data.

Chapter 4 deals with the technical implementation of the methods described in the previous section. The implementation required the development of a computer program. Computationally, there are numerous considerations in order to successfully apply these methods in an epidemiological context, which are discussed here. This section also provides experimental analysis of computational performance, and technical details of the program.

The next three chapters (5, 6, and 7), are the results chapters and represent the bulk of the intellectual effort of the thesis. Each of these chapters applies Bayesian networks to Health Surveys for England data to tackle a problem in obesity epi-

demiology.

Chapter 8 is the final chapter of this thesis; it provides a discussion of the results obtained in the previous three chapters. The chapter also seeks to draw together some of the themes explored and comments on the utility of the Bayesian structural modelling in an epidemiological context. The shortcomings and potential future development of this approach is considered.

Chapter 2

Data

This chapter describes the datasets used in this thesis. The primary source of data are the Health Surveys for England (HSE), data from the 2001 UK census is also employed. Background information is provided on both datasets, including method of collection, number of participants as well as limitations of the data. Details and definitions of the variables used throughout this thesis are also included.

2.1 Health Surveys for England Data

2.1.1 Overview

The HSE are a series of annual health surveys sponsored by the Department of Health (`http://www.doh.org`) that have been conducted since 1991 to determine the overall health of people living in England. This chapter provides a summary of all variables used in the three results chapters of this thesis. Within each chapter variable selection was performed separately depending on the research question.

The majority of HSE data is derived from questionnaire based interviews conducted by a trained interviewer at the participating household. The HSE also includes the collection of physical measurements and blood samples for analysis by a visiting nurse. Additional information is obtained from self completion booklets. Detail of how each data from each section of the HSEs was collected is provided by the tables in appendix B.1. Briefly, the majority of data including socio demographic characteristics and physical activity behaviour were derived from interview questions. Physical measurements such as weight and height data were collected by a trained nurse. Dietary intake information was collected by self completion booklets, a format that is known to have issues with bias. Individuals of all ages are surveyed, though different data are collected from children.

Data are sampled from households rather than individuals and are chosen randomly by postcode. Following invitation, households that decline to participate are not replaced. Participation bias is therefore likely [129], but there are no data available regarding participation rates between areas [Personal Communication-UKDA]. Data are freely available to download from the UK Data Archive (UKDA)[1] subject to registration.

The HSE surveys consist of a core and extended element. The core element consists of standard questions that are consistent between years. The extra component varies to focus on a particular disease or population group. In 2003 and

[1] `https://www.data-archive.ac.uk/findingData/hseTitles.asp`

2006 the HSE focused on cardiovascular disease (CVD) and associated risk factors. Supplementary questions asked for detailed information on CVD symptom history and relevant risk factors such as physical activity, diet, and smoking habits; hence providing information relevant to obesity.

The 2003 survey selected 13,680 addresses, 19 each from 720 postcode sectors. Interviews were conducted at 73% of selected households, 90% of adults within cooperating households participated. In total data was collected for 14,836 adults and 3,717 children. The 2006 survey included a general population sample of adults and children, as well as a boost sample of 2-15 year olds. In the general sample 14,400 addresses were selected from the 720 postcode sectors. Excluding the boost sample, interviews were conducted at 63% of households, at which 88% of adults participated. The final sample contained 14,142 adults and 7,257 children.

In the 2006 survey individuals were assigned to one of two sample groups (CORE 1 and 2). In order to reduce load on older participants these groups were given abbreviated questions on specific topics if over the age of 65. Individuals over 65 in the CORE 2 group were not asked questions from the cardio-vascular disease section of the survey, this section included detailed questions on medical history of CVD and diabetes. Instead, this group were asked questions from the long version of the physical activity questionnaire, while the CORE 1 group completed the attenuated version. Consequently, individuals over 65 in the CORE 1 group have insufficient data collected to inform several of the energy expenditure variables used. Irrespective of sample group, individuals under 65 were asked to complete all questions. Variables from the HSE are held consistent between years where possible, to ease comparability between surveys. Variables described are available in both 2003 and 2006 datasets, unless otherwise specified.

Due to constraints of the Bayesian network method (see chapter 3), it is necessary to use categorical variables in these analyses, this unavoidably results in a loss of precision from continuous variables. Although mixed discrete-continuous models are possible, they require approximation methods which add unmanageable complexity. This is discussed in more detail in Chapter 8.

2.1.2 List of Variables

Variables derived from HSE data are listed below, with a brief description and the number of categories (n).

Socio-Demographic Variables		
Variable	**Description**	***n***
Sex	Declared gender	2
Age	Age groups	6
Dependent Children	Whether children in household to which individual is a parent or step-parent	2
Marital Status	Whether in relationship with another individual living in same home	2
Health status	Self-reported general health	3
NS-SEC	Socio-economic classification	5
Economic Activity	Economic position last week	4
Ethnicity	Ethnic group	2
Education	Highest educational qualification	4
Leisure Access	Whether believes access to leisure facilities is poor	2
Transport Access	Whether believes access to public transport is poor	2

Energy Intake Variables		
Variable	**Description**	***n***
Fried food intake	Weekly intake (groups)	3
Cake/sweets intake	Weekly intake (groups)	4
Snack/crisps intake	Weekly intake (groups)	4
Fruit/vegetable intake	Weekly intake (groups)	4
Energy Expenditure Variables		
Variable	**Description**	***n***
Recreational activity	Time spent participating in recreational physical activity per week (groups)	4
Incidental activity	Time spent walking (groups)	3/4 [1]
Occupational activity	Work/daytime activity level (groups)	4

[1] Questions related to walking are not consistent between 2003 and 2006, the variables are not equivalent. Consequently, different variables were generated. See next section for details.

Waist Hip Ratio Variables		
Variable	**Description**	***n***
BMI	BMI groups	6
WHR	WHR groups	6
Alcohol intake	Weekly consumption (grouped)	3
Smoking status	Current smoking behaviour	3
Period status	Whether still menstruating (females only)	2

Socio-Demographic Variables (matched with 2001 census) [1]		
Variable	**Description**	***n***
Age*	Age groups	6
Social Status	National Readership Survey groups	5
Economic Activity*	Economic position last week	4

[1] Different groupings to above variables to allow matching with census data. Variable names of these alternative groupings are distinguished with an asterisk (*).

2.1.3 Description of Variables

This section provides detailed information on the derivation of the variables used in this thesis. If a variable is *Derived* it indicates that it is made up from a combination of HSE variables, details of which are provided. *Grouped* indicates that the variable is based on a named HSE variable, but categories have been re-evaluated. If taken directly from HSE data, the variable is labelled *Original*. Precursor variables from the original HSE dataset are listed. Precise derivations of variables are included in the appendix as STATA[2] code.

Socio-Demographic Variables

SEX

Status: Original.
Precursors: SEX.
Groups: Male, Female.

[2] www.stata.com

Notes: Directly taken from HSE data.

AGE

Status: Grouped.
Precursors: AGE.
Groups: 16-24, 25-34, 35-44, 45-54, 55-64, 60-74.
Notes: Age categories. Generated from age at last birthday variable in HSE.

DEPENDENT CHILDREN

Status: Derived.
Precursors: PSERIAL, HSERIAL, RELTO01-RELTO12.
Groups: Yes, No.
Notes: Whether an individual is a parent or step-parent to one or more children living in the household. NB: This measure may miss young mothers living with own parents.

MARITAL STATUS

Status: Derived.
Precursors: MARITAL, COUPLE.
Groups: Single, Cohabiting.
Notes: Whether an individual is living with a partner.

HEALTH STATUS

Status: Original.
Precursors: GENHELF2.
Groups: Good, Fair, Poor.
Notes: General health. Directly from HSE dataset, response to question 'How is your health in general?'.

NSSEC (SOCIAL CLASS)

Status: Derived.
Precursors: HPNSSEC8, NSSEC8, NSSEC3.
Groups: Managerial/Professional, Intermediate, Routine/Manual, Long-term Unemployed, Other (not classified).
Notes: Three category version of the National Statistics Socio-Economic Classification. Taken from the NSSEC of the family reference person if available, otherwise of individual.

Economic Activity

Status: Grouped.
Precursors: ACTIVB,TOPQUAL2.
Groups: Employed/Student, Unemployed, Retired, Looking after home or family.
Notes: Economic activity last week.

Ethnicity

Status: Grouped.
Precursors: ETHINDA.
Groups: White, Non-White.
Notes: Ethnicity of individual. Due to small numbers, non white groups merged.

Education Level

Status: Derived.
Precursors: TOPQUAL2, ACTIVB.
Groups: UK Higher Education, Below Higher Education, No Qualifications, Current Student.
Notes: Highest qualification attained. Those with non-UK qualifications placed in 'Below Higher Education' category.

Leisure Access

Status: Grouped.
Precursors: LEISURE.
Groups: Poor, Fair/Good.
Notes: Response to statement 'This area has good leisure things for people like myself?'. Four possible responses; strongly agree, agree, disagree, strongly disagree. Strongly disagree is coded as 'Poor' otherwise 'Fair/Good'. This coding was decided upon following tabulation of the variable, 'poor' was the most common response, and I wanted this variable to identify individuals in very badly served areas.

Transport Access

Status: Grouped.
Precursors: TRANSPRT.
Groups: Poor, Fair/Good.

Notes: Response to statement 'This area has good local transport'. Four possible responses; strongly agree, agree, disagree, strongly disagree. Strongly disagree is coded as 'Poor' otherwise 'Fair/Good'. This coding was decided upon following tabulation of the variable, 'poor' was again the most common response, and I wanted this variable to identify individuals in very badly served areas.

Energy Expenditure Variables

RECREATIONAL PHYSICAL ACTIVITY

Status: Grouped.

Precursors: HRSSPTG.

Groups: None, 0-1, 1-3, 3+.

Notes: Hours of recreational physical activity per week, calculated from hours performed in last 28 days. Includes active sports, running, swimming, cycling, but not activities such as golf, hiking etc. For full detail see HSE documentation [130].

INCIDENTAL PHYSICAL ACTIVITY- 2003 VERSION

Status: Derived.

Precursors: DAYWLK30, WLK5INT, WLK30.

Groups: No walks of at least 5 minutes,No walks of at least 30 minutes, walks over 30 minutes (in the last month).

Notes: Walking behaviour in last month. Questionnaire has different structure than 2006, less detail, variables not equivalent.

INCIDENTAL PHYSICAL ACTIVITY- 2006 VERSION

Status: Derived.

Precursors: DAYWLK, WLK5INT, WLK15.

Groups: <2, 2-10, 10-20, 20+..

Notes: Number of walks greater than 15 minutes in the last month.

OCCUPATIONAL PHYSICAL ACTIVITY

Status: Derived.

Precursors: WORKACT, HWRKLIST, GARDLIST, HEVYH, MANWORK.

Groups: Inactive, Low Activity, Moderate, Active.

Notes: Nominal scale, based on an additive combination of variables that describe physical activity done at work, and during housework, DIY, and gardening. See derivation in appendix for full detail.

Energy Intake Variables

Fried Food Intake

Status: Grouped.
Precursors: FRIEDFDB.
Groups: <1, 1-2, 3+.
Notes: Number of times fried food eaten per week.

Snack Intake

Status: Grouped.
Precursors: SNACK.
Groups: <1, 1-2, 3-5, 5+.
Notes: Number of times snacks, crisps, biscuits, nuts etc. eaten per week. Although nuts are high in protein and low in saturated fats, they are included in this particular HSE question. This is not desirable, and would be better excluded, or more detail provided..

Cake Intake

Status: Grouped.
Precursors: CAKESC.
Groups: <1, 1-2, 3-5, 5+.
Notes: Number of times cakes or sweets etc. eaten per week.

Fruit and Vegetable Intake

Status: Grouped.
Precursors: PORFTVG.
Groups: ≤1, 1-2, 3-5, 5+.
Notes: Number of portions of fruits or vegetables eaten per day.

Waist-Hip Ratio Variables

Body Mass Index

Status: Original.

Precursors: BMIVG6.
Groups: ≤19, 19-24.9, 25-29.9, 30-34.9, 35-39.9, 40+.
Notes: Body Mass Index (BMI) groups based on WHO classifications [1].

WAIST HIP RATIO

Status: Original.
Precursors: MENWHGP, WOMWHGP.
Groups: males: ≤0.80, 0.80-0.85, 0.85-0.90, 0.90-0.95, 0.95-1.00, 1.00+; females: ≤0.70, 0.70-0.75, 0.75-0.80, 0.80-0.85, 0.85-0.90, 0.90+.
Notes: Waist Hip Ratio (WHR) groups, boundaries differ by gender.

ALCOHOL INTAKE

Status: Grouped.
Precursors: DNOFT2.
Groups: <1 (light), 1-4 (moderate), 5+ (heavy).
Notes: Number of days on which alcohol consumed per week.

SMOKING STATUS

Status: Original.
Precursors: CIGSTA3.
Groups: Current cigarette smoker, Ex-regular cigarette smoker, Never regular cigarette smoker.
Notes: Current smoking behaviour.

PERIOD STATUS

Status: Original.
Precursors: PERIOD.
Groups: Yes, No.
Notes: Females only: Response to question 'Are you still menstruating?'.

Variables for 2001 Census

AGE*

Status: Grouped.
Precursors: AGE.
Groups: 16-19, 20-29, 30-39, 40-49, 50-64, 65-74.

Notes: Generated from AGE variable in HSE. Categories chosen to allow matching with 2001 census data.

ECONOMIC ACTIVITY*

Status: Grouped.
Precursors: ECONA.
Groups: Employed, Unemployed, Other economically inactive, Student.
Notes: Categories chosen to match with 2001 census. Data on whether retired or a homemaker not available in census, consequently these categories coded here as 'Other inactive'.

SOCIAL STATUS

Status: Original.
Precursors: SCHRPG4.
Groups: Managerial, Manual, Semi/Un-skilled, Unemployed, Unknown.
Notes: Variable chosen to allow comparability with 2001 census. In 2001 census data NSSEC not available if retired or current student. Categories from National Readers Survey (NRS) classification.

2.1.4 Strengths and Weaknesses of Data

The HSEs are a valuable source of information for researchers collecting extremely detailed health information from a large number of people. Nurse visits ensure medical readings are as accurate as possible. However, participation bias is likely within the HSE, due largely to non-response [129]. The HSE is likely to underrepresent individuals of lower social class, non-native English speakers and young people living away from home [129].

Much of HSE data is self-reported and therefore subject to bias [131, 132]. Obese individuals are known to underreport dietary intake [131], and self reported physical activity levels are also known to often be innaccurate [133]. Direct measures of energy expenditure, such as an accelerometer and doubly labelled water are available, but expensive and impractical for large studies. Validation of HSE physical activity has been carried out for children [134], and has been labelled 'unreliable'. In this thesis I did not use absolute levels of physical activity, but rather grouped the population into categories. This ensures, barring a systemic bias, that individuals are in a group that is true to their genuine behaviour.

2.2 UK Census 2001

2.2.1 Overview

Chapter 6 applies data from the 2001 UK Census to a model of health behaviour learnt from HSE data. The organisation of the 2001 UK census was overseen by the Office of National Statistics (ONS). Censuses in the UK are carried out every ten years, with the intention of obtaining demographic and employment information over the whole population. Self completion forms were delivered to households and communal establishments three weeks before the census date of the 29^{th} April 2001. Completed censuses were returned by post.

The postal response rate was 88% for the England and Wales [135], with follow up by enumerators if forms were not returned. Overall, it was estimated that 94% individuals living in England and Wales were recorded in the census [135]. Various reasons explain how some individuals were not recorded [136].

A number of data products are available for census data. Summary statistics over geographically small units (census output areas) are available, however, individual level data is more restricted and only provides relatively course geographical information. The Samples of Anonymised Records (SARs) are a set of datasets providing a random sample of individuals or households from census data. There are two datasets that provide individual level samples from 2001 Census data; The 2001 Individual Licensed SAR (IL-SAR), this is a 3% sample containing information on a full range of census topics, the smallest geographical unit available is the Government Office Region (GOR); the Small Area Microdata (SAM) is a 5% sample with less individual detail, but with a finer geographical resolution (Local Authorities). In this application I use the SAM dataset, mainly due to this finer geographical referencing.

Variables were selected with the intention of providing a good representation of socio-demographic factors, that also had equivalent variables in the HSE data. The presence of the letter ‘c’ in the variable name distinguishes Census variables from HSE variables.

2.2.2 List of Variables

2001 Census Variables		
Variable	**Description**	***n***
cSex	Sex	2
cAge	Age groups	6
cDependent Children	Number of children in household to which individual is a parent or step-parent	2
cMarital Status	Whether in relationship with another individual living in same home	2
cHealth status	Self judged general health	3
cSocial Status	National Readership Survey (NRS) social goups	5
cEconomic Activity	Economic position last week	4
cEthnicity	Ethnic group	2
cEducation	Highest educational qualification	4

2.2.3 Description of Variables

This section provides a detailed definition of the variables used. Precursor variables are those in the 2001 SAM data from which the variables here were derived. Full STATA code of variable derivations is provided in the appendix.

CSEX

Status: Orignal.
Precursors: SEX.
Groups: Male, Female.
Notes: As original.

CAGE

Status: Grouped.
Precursors: AGE.
Groups: 16-19, 20-29, 30-39, 40-49, 50-64, 65-74.
Notes: Age variable in census data categorised, hence need to re-evaluate HSE groups to allow matching.

CHEALTH

Status: Original.
Precursors: HEALTH.
Groups: Good, Fair, Poor.
Notes: Response to ‘General health in the last 12 months?’.

cDependent Children

Status: Derived.

Precursors: FNDEPCHA, RELTO, GEN.

Groups: Yes, No.

Notes: Whether individual parent or step parent to one or more children living in household. NB: This measure may miss young mothers living with own parents.

cMarital Status

Status: Derived.

Precursors: FAMTYP, RELTO, GEN.

Groups: Single, Couple.

Notes: Whether in relationship with another person living at address.

cEconomic Activity

Status: Grouped.

Precursors: ECONA.

Groups: Employed/Student, Unemployed/Looking for work, Other economically inactive.

Notes: Data on whether retired or a homemaker were not available. Matched with Economic Activity*.

cSocial Status

Status: Original.

Precursors: HRSOC.

Groups: Managerial, Manual, semi/un-skilled, Unemployed, Unknown.

Notes: In 2001 census data NSSEC not available if retired or current student, so social status used. Categories taken from NRS (National Readership Survey).

cEducation Level

Status: Derived.

Precursors: QUALV, STUDENT.

Groups: UK Higher Education, Below Higher Education, No Qualifications, Current Student.

Notes: Highest qualification attained. Those with non-UK qualifications placed in 'Below Higher Education' category.

CETHNICITY

Status: Grouped.
Precursors: ETHEWA.
Groups: White, Non-white.
Notes: Ethnicity. Non white individuals merged into single group.

2.2.4 Strengths and Weaknesses of Data

The UK census is a mandatory survey, and thus has a very high participation rate of circa 94%. Questions are formal and closely defined, with little scope for bias. Although the remit of the census is very broad, the information provided in terms of deprivation is very accurate with several useful proxy indicators such as vehicle ownership, house ownership status, people/room etc. The census data used here was carried out in 2001, which may mean various boundaries and population statistics are outdated. Nevertheless, the 2001 census represents the best available source for demographic information on the UK population.

2.3 Discordance between HSE 2006 and 2001 Census

Chapter 6 uses variable matching between HSE and Census data, the following table (2.1) summarises differences in these variables between datasets.

	2001 Census: Self report form	2006 HSE: Nurse Administered Questionnaire
Age	What is your date of birth?	Can I check your age last birthday?
Sex	What is your sex? A Male, B Female.	INTERVIEWER: CODE (name of respondents) SEX.
Dependent Child	Adult living in household parent or step-parent to a dependent child. A dependent child is a person aged 0 to 15 in a household (whether or not in a family) or aged 16 to 18 in full-time education and living in a family with his or her parent(s).	Whether individual is a parent or step parent of ≤18 year old in household.
Marital Status	Whether another person living at same address listed as partner or 'spouse.	Response to 'Are you living with anyone in this house?'
Health Status	How would you describe your general heath over the last 12 months? Good/Fair/Poor	How is your general heath? Good/-Fair/Poor
Social Class	Social grade is a socio-economic classification used by the Market Research and Marketing Industries. The algorithm for deriving Approximated Social Grade was developed with the Market Research Society. Of household reference person.	Social class of household reference person. Based on classification of occupation as provided by respondent.
Economic Activity	Last week, were you doing any work: as an employee or on a Government sponsored training scheme, as self-employed/freelance, or in your own family business? If not, were you any of the following? Retired, Student, Looking after home/family, Permanently sick/disabled?	Which of these descriptions (selection) applies to what you/-name (Household Reference Person) were doing last week, that is in the seven days ending (date last Sunday)?
Education Level	Which of these qualifications do you have? (selection)	Do you have any of the qualifications listed on this card? Please look down the whole list before telling me. (more detailed selection).
Ethnicity	What is your ethnic group? A White, B Mixed, C Asian or Asian British, D Black or Black British, E Chinese or Other ethnic group.	Can I check, to which of the groups on this card do you consider you belong? A White, B Mixed, C Asian or Asian British, D Black or Black British, E Chinese or Other ethnic group.

Table 2.1: Potential discord between matched variables in the 2006 HSE and 2001 census

Chapter 3

Methodology

This chapter provides a detailed description of the methods implemented in this thesis. Firstly, a brief overview of Bayesian networks is provided, before moving on to a more general discussion of Bayesian statistics. From this groundingI proceed to a more complete specification of Bayesian networks and introduce methods for learning Bayesian networks from data. The chapter then outlines the main application of Bayesian networks in this thesis; how their structure can help to identify dependency relations between variables within datasets.

3.1 Introduction to Bayesian Networks

3.1.1 Overview

A Bayesian network (BN) is a graphical representation of a joint probability distribution. A BN is composed of nodes, each representing a random variable; and arcs, each representing a conditional dependency between two variables. The structure of the network is therefore simply a set of assertions of conditional dependence within a set of random variables.

A Bayesian network takes the form of a Directed Acyclic Graph (DAG), and consists of two components; the network structure (or *topology*) designated H, and parameters θ; that specify the precise probabilistic relations between variables. An example of a simple Bayesian network including probability tables is shown in figure 3.1 (image has been released into the public domain, and does not represent author's own work).

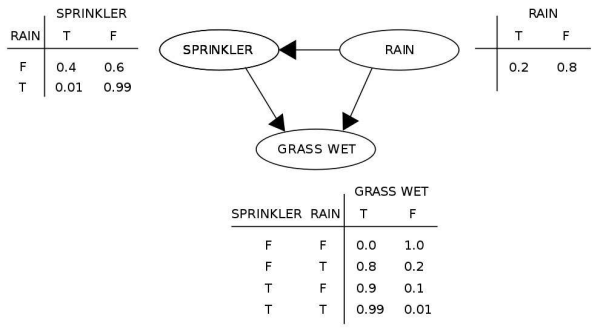


Figure 3.1: Example of a simple Bayesian network representing influence of *'Rain'* and *'Sprinkler'* on *'Wet Grass'*

In this example the state of *'Wet Grass'* is conditionally dependent on the other

two variables *'Rain'* and *'Sprinkler'*. Additionally *'Sprinkler'* is conditionally dependent on *'Rain'*. Each of these three variables has two states; True (t) and False (f). The example network shown is equivalent to the joint probability function in eq. 3.1 (variable names *'Grass Wet'*,*'Sprinkler'* and *'Rain'* are abbreviated to *G*, *S* and *R* respectively).

$$\Pr(G,S,R) = \Pr(G|S,R)\Pr(S|R)\Pr(R). \tag{3.1}$$

This model enables the calculation of conditional probabilities, such as $\Pr(G|R = t)$ by *marginalising* over variables. When networks become large, the necessary calculations become exponentially more complex, and approximate techniques must be used.

Bayesian Network models can be applied to classification tasks, for example in medical diagnostics [108, 137] and evaluating economic risk [110, 138]. A Bayesian network model can be constructed from expert opinion, learned from data, or a combination of both. A key feature of Bayesian networks is the ability to calculate the *likelihood* of the observed data given it was generated from a particular network. This likelihood score can then be used to form a measure of the evidence associated with that particular network structure. Section 3.3 discusses the learning of network structure and parameters from data. As noted earlier, the structure of a Bayesian network represents a set of assertions of conditional dependencies within data. By learning network topology from data it is possible to identify these conditional dependencies without investigator intervention.

3.1.2 D-Separation

Conditional independence is an important concept when applying Bayesian networks. If two nodes have no path between them we can say that they are *d*-separated, this means that knowledge of the state of one provides no information of the state of the other. This is defined explicitly in Pearl (1988) [139]. Two sets of nodes *X* any *Y* are *d*-separated if and only if every path between them is blocked. As illustrated in figure (3.2), blocked in this context means an intermediate variable *V* exists between *X* and *Y*. *X* and *Y* are *d*-separated if:

- *X* and *Y* are connected serially with *V* between them (3.2(a)), *V* is instantiated (known).
- *X* and *Y* diverge from *V* (3.2(b)) which is instantiated.

- X and Y converge to V (3.2(c)), and V nor any of $V's$ descendants is NOT instantiated.

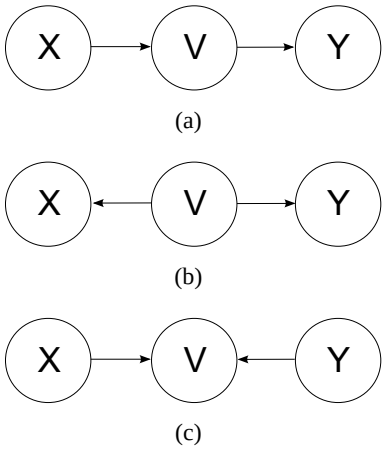


Figure 3.2: Diagram showing conditions of d-separation in Bayesian networks

A specific network structure H allows us to distinguish between correlation and direct dependence [124]. Two variables X and Y may be correlated, however if the state of a third variable V is known X and Y may be independently distributed (*i.e.* d-separated). The relationship between X and Y is said to be mediated by the state of V. This feature allows the use of BNs to identify the key variables in complex multivariate probability distributions.

The *Markov blanket* of a node is a term used to describe the minimum set of variables that can d-separate the node from the rest of the network. The Markov blanket of node X consists of $X's$ parents, children, and parents of its children.

3.1.3 Causality

Although BNs are often used to describe causal relationships between variables (e.g. figure 3.1), the presence of an arc $X \rightarrow Y$ does not require or imply a causal relationship between X and Y. The existence of an arc merely indicates that the probability distribution of the state of Y if influenced by the state of X. Equally, if Y is known this provides us with information regarding the likely state of X; consequently $Y \rightarrow X$ is an equivalent representation of the conditional dependency relationship between the two variables. When Bayesian networks are learned from data, an assumption that conditional dependencies are causal can be highly misleading.

3.2 Bayesian Approach to Statistics

A Bayesian Network is a joint distribution of random variables. Although Bayesian networks can model continuous data, in this application each variable is discrete with finite outcomes. In the current section we examine a single variable in isolation from the rest of the network, networks involving several variables are discussed later. A more thorough grounding is available elsewhere [140, 141].

Let X represent a random multinomial variable with states $A = \{a_1....a_r\}$. Bayesian statistics allows us to update our prior beliefs following observed data. The data D is in the form $\{X_1 = x_1...X_N = x_N\}$, $x_1....x_n \in A$, and can be summarised by the counts $\{n_{x_1}....n_{x_r}\}$. The quantity we wish to evaluate is the probability of that the next observation is a_k given data already observed $\Pr(X = a_k|D)$.

The parameters θ specify the probability of each possible state of X. Here, the model is parameterised by a probability vector that is normalised to sum to 1. θ_k is the probability that an observation of X will be in state a_k. In Bayesian statistics the *posterior* distribution of θ is derived from the addition of data to the *prior* distribution of θ. The posterior probability of θ ($Pr(\theta|D)$) is obtained from the prior of θ ($Pr(\theta)$) using Bayes' rule.

$$\Pr(\theta|D) = \frac{\Pr(\theta)\Pr(D|\theta)}{\Pr(D)}. \tag{3.2}$$

The marginal likelihood of the observed data, $\Pr(D)$ is invariant with θ, and can be disregarded in this instance.

$$\Pr(D) = \int \Pr(D|\theta)\Pr(\theta)d\theta \tag{3.3}$$

The term $\Pr(D|\theta)$ is the likelihood function; the probability of the observed data given θ. This value, assuming independent observations, is simply the product of θ_k for each observation:

$$\Pr(D|\theta) = \prod_{k=1}^{r} \theta_k^{n_k} \tag{3.4}$$

Incorporating the likelihood function into eq. 3.2 the posterior distribution becomes:

$$\Pr(\theta|D) = \frac{\Pr(\theta)\prod_{k=1}^{r}\theta_k^{n_k}}{\Pr(D)} \tag{3.5}$$

$\Pr(X = a_k|D)$ is equivalent to the expectation of θ_k. The expectation of θ is calculated by the integration of the *posterior* distribution (3.5) with respect to θ. How-

ever, before integration is possible we must assign a distribution to the *prior* $\Pr(\theta)$. For convenience a *Dirichlet* distribution is assumed (eq. 3.7. The Dirichlet distribution is a continuous multivariate probability distribution. The probability density function describes the possible values of the vector of probabilities θ associated with r rival events. Each component of the r-dimensional vector is in the range $[0, 1]$, and the vector must sum to 1. The parameters of the Dirichlet (α) are r positive real numbers. The density function shown in eq. 3.7 returns the probability of the vector θ given that each of the r possible events has been observed $\alpha - 1$ times:

$$\Pr(\theta) \equiv \frac{\Gamma(\alpha)}{\prod_{k=1}^{r} \Gamma(\alpha_k)} \prod_{k=1}^{r} \theta_k^{\alpha_k - 1}. \tag{3.6}$$
$$\text{where } \alpha = \sum_{i=1}^{r} \alpha_k \text{ and } \alpha_k > 0.$$

These α values are known as the *hyperparameters* or *pseudocounts*. These are assigned by the experimenter and represent prior knowledge. Various conventions and techniques exist for assigning priors [142]. Priors can either be set to be uninformative, or reflect some degree of certainty in the distribution of θ. The chief advantage of specifying a Dirichlet prior is that the resulting posterior distribution follows a Dirichlet distribution:

$$\Pr(\theta|D) \equiv \frac{\Gamma(\alpha + N)}{\prod_{k=1}^{r} \Gamma(\alpha_k + n_k)} \prod_{k=1}^{r} \theta_k^{\alpha_k + n_k - 1} \tag{3.7}$$

Owing to this convenience it is said that the prior and posterior are *conjugate*. The expectation of θ_k is therefore straightforward to compute:

$$\Pr(X_{N+1} = x_k|D) = \int \theta_k \text{Dir}(\theta|\alpha_1 + n_1, ..., \alpha_r + n_r) d\theta = \frac{\alpha_k + n_k}{\alpha + N} \tag{3.8}$$

This expecting reflects the probability that the next observation of variable X will be in state a_k. The expectation is the integral of posterior distribution with respect to θ. The influence of the prior on the final expectation of θ_k can be seen in equation 3.8, as the observed counts (n_k) grow larger the influence of the pseudocounts (α_k) diminishes.

3.3 Learning of Bayesian Networks

3.3.1 Learning Network Structure

The methods explored in this thesis involve the discovery of graph structure from data; this requires the evaluation of the probability of a network topology (H) given the observed data (D), expressed as $\Pr(H|D)$. Further detail of learning structure of Bayesian networks is available more fully elsewhere [124, 140], an overview is provided below.

A Bayesian network represents a joint distribution of a set of randomly distributed variables $\mathbf{X} = \{X_1, ...X_n\}$, where **boldface** denotes a set of variables or configurations of variables. The set of parents of a given variable X_i are denoted π_i. A node Y is a parent of X if there is a directed arc from $Y \rightarrow X$, which specifies a conditional dependency. Different network topologies (H) specify different joint distributions:

$$\Pr(\mathbf{x}) = \prod_{i=1}^{n} \Pr(x_i|\pi_i) \tag{3.9}$$

The parentset π_i has a number of possible configurations $B = \{b_i, ..., b_q\}$ equal to the product of the number of levels of each member of the set. Each state b_j of π_i corresponds to an input level of X_i. The parameters of a specified network (denoted as θ_H) contain a vector θ_{ij} for each node i, and input j combination; the distribution of which can be represented by a Dirichlet as described in the previous section.

If the probability of topology given data, $\Pr(H|D)$, is the quantity we wish to evaluate, simple application of Bayes' Rule provides:

$$Pr(H|D) = \frac{\Pr(D|H)\Pr(H)}{\Pr(D)}. \tag{3.10}$$

Equation 3.3 provided the marginal likelihood, $\Pr(D)$, of a simple one variable model. By substituting the likelihood function and prior probability in eq. 3.3 with equations 3.4 and 3.7, we derive equation 3.11. This quantity represents the probability of the observed data marginalised over possible values of θ:

$$\begin{aligned} \Pr(D) &= \int \frac{\Gamma(\alpha)}{\prod_{k=1}^{r}\Gamma(\alpha_k)} \prod_{k=1}^{r} \theta_k^{\alpha_k - 1} \prod_{k=1}^{r} \theta_k^{n_k} d\theta \\ &= \frac{\Gamma(\alpha)}{\prod_{k=1}^{r}\Gamma(\alpha_k)} \int \prod_{k=1}^{r} \theta_k^{\alpha_k + n_k - 1} d\theta \end{aligned}$$

$$= \frac{\Gamma(\alpha)}{\prod_{k=1}^{r}\Gamma(\alpha_k)}.\frac{\prod_{k=1}^{r}\Gamma(n_k+\alpha_k)}{\Gamma(N+\alpha)}\int\frac{\Gamma(N+\alpha)}{\prod_{k=1}^{r}\Gamma(n_k+\alpha_k)}\prod_{k=1}^{r}\theta_k^{\alpha_k+n_k-1}d\theta$$

$$= \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}\prod_{k=1}^{r}\frac{\Gamma(\alpha_k+n_k)}{\Gamma(\alpha_K)} \tag{3.11}$$

This quantity refers to the single variable case. This marginal likelihood expression can be extended to a joint distribution of variables by taking the product of the marginal likelihood for each θ_{ij} in the network to provide the likelihood of data given topology, Pr($D|H$), *independently* of parameters. This is shown in equation 3.12:

$$\Pr(D|H) = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})}.\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})} \tag{3.12}$$

This equation provides the Pr($D|H$) term in equation 3.10. The Pr(D) term in this equation reflects the probability of the data, over all θ and H. In the single variable model there is no concept of topology, so the Pr(D) in equation 3.11 has not been marginalised over all topologies, and is not equivalent to that in equation 3.10. The Pr(H) term is the prior probability of the network topology. In this thesis I use an uninformative complexity penalising prior, as implemented in [143]:

$$\Pr(H) = \frac{1}{\Pi}\prod_{n=1}^{N}\binom{N-1}{|\pi_n|}^{-1} \tag{3.13}$$

The cardinality of the parent set of n is expressed as $|\pi_n|$ and $\frac{1}{\Pi}$ is a normalisation constant. The methodology used makes the normalisation constant irrelevant and hence it is not specified, this is explained later. Many alternative strategies exist for specifying a prior over network structures, such as using a uniform prior, or penalising networks based on some distance measure [140] from some pre-specified structure. Other approaches are available [142, 144, 145].

The probability of the data, Pr(D) is invariant with structure, as it represents the marginal probability over topologies and parameters and can be disregarded. Following this, according to Bayes' theorem (3.10), the product of the Pr($D|H$) and Pr(H) (equations 3.12 and 3.13) are directly proportional to Pr($H|D$):

$$\Pr(H|D) \propto \Pr(D|H)\Pr(H) \tag{3.14}$$

This forms the criterion used to score network topologies used throughout this

thesis. This a measure of the evidence of a specified network structure given the observed data; or more accurately, the likelihood of the observed data given the network topology in addition to a prior over network structures. This measure is not the posterior probability of the network structure, but as we have seen, it is directly proportional to it. This criterion can be used to find the most probable network given data, or to sample over possible network topologies to estimate the integral or expectation of the network topology distribution.

3.3.2 Learning Network Parameters

Given network structure, the parameters of the network need to be assigned. Each node i of a complete Bayesian network possesses a probability vector for each state k given each combination of input states j. This vector of length r is θ_{ij}. Within a network, there are likely to be a significant number of such probability vectors, depending on the network topology H. Optimisation over both topology and parameters is difficult owing to changes in topology resulting in a completely distinct probability landscape for parameters. Each θ_{ij} is distributed independently, thus each can be evaluated separately.

The vector θ_{ij} represents the parameters of a multinomial distribution with r possible states. As the probability vector described in section 3.2, the prior of θ_{ij} is assumed to be Dirichlet distributed, resulting in the following posterior following incorporation of data:

$$\Pr(\theta_{ij}|D) \equiv \frac{\Gamma(\alpha + N)}{\prod_{k=1}^{r} \Gamma(\alpha_k + n_k)} \prod_{k=1}^{r} \theta_{ijk}^{\alpha_k + n_k - 1} \tag{3.15}$$

The parameters of the above Dirichlet distribution are the sum of the prior and the counts of the observations in each output level of node i given input j. The simplest method of assigning parameters to a network is to take the value of θ_{ij} that maximises this probability. However, this approach does not take into account the uncertainty associated with fewer observations, the greater the number of observations the more reasonable this approach is as data overwhelms the prior.

3.3.3 Assumptions of Bayesian Networks

The use of Bayesian networks in this context are subject to several assumptions.

- **Independence of observations**. All datapoints are assumed to be independent, this is necessary for the straightforward calculation of likelihood in

eq. 3.4. If this assumption were not met, the likelihood function would be extremely difficult to evaluate.

- **Dirichlet distributed prior**. Although it possible to use distributions other than the Dirichlet as the prior, its choice allows easy integration of the posterior provability distribution as shown in eq. 3.8.

- **Discrete data**. Bayesian networks may be used to model a set of continuous random variables. Mixing of discrete and continous data in Bayesian networks is extremely complicated due to the non conjugate nature of the respective distributions. Although some of the variables included in the models described here are naturally continuous, they have been discretised for convenience. The alternative would require computationally expensive approximation techniques.

- **Non missing data**. The models implemented here cannot incorporate missing data. Although there are numerous methods available to calculate the evidence for network structures despite missing data, these rely on approximations and are computationally intensive [146, 147]. An alternative is to include missing values as a separate category in each variable, however in this application would result in arcs explaining missing data rather than true interdependencies.

3.4 Bayesian Model Averaging

A single network structure H, represents a number of conditional dependencies between variables. The structure of a network therefore provides information regarding conditional dependencies within a dataset. The crux of the work in this thesis applies this property of Bayesian networks to reveal information about codependencies in complex datasets. However, given data D there are a huge number of possible distinct network structures that could explain the observed data. The relative probability that each network generated the observed data can be calculated (section 3.3.1). Although there is likely to be a single most probable network, this may only represent a tiny proportion of the total probability integral over the space of all possible networks. Consequently, the structure of the highest scoring network is not necessarily a good indicator of relationships present.

In order to include a contribution from all possible network structures, Bayesian Model Averaging (BMA) is applied, as described in Madigan and York [144]. The

central concept of BMA is to identify structural features present in the majority of the total probability integral of all networks. These features provide an estimate of the conditional dependencies present in the data. A structural feature may be an arc between two nodes, or the presence of a node within the Markov blanket of another [124].

To calculate the posterior probability of the existence of various structural features, the probability density of the topologies in which a feature is present relative to the entire space of possible topologies G is estimated. Evaluation of the entire space of Bayesian networks is infeasible due to its magnitude. The exact number of possible networks a_n depends on the number of nodes, n, and can be calculated using the recurrence relation [148]:

$$a_n = \sum_{k=1}^{n} (-1)^{k-1} \binom{n}{k} 2^{k(n-k)} a_{n-k} \tag{3.16}$$

The most straightforward approximation to the space G is the selection of the most probable network. In a domain with infinite data, the probability distribution of the topology space becomes a Dirac-delta function, *i.e.* one topology becomes infinitely more probable than the others. As the number of observations increases this approach becomes more reasonable. A much superior approach is to average over a number of Bayesian network structures. A set $\hat{G}$ of high scoring topologies is derived, forming an approximation to the entire space G [144,149]. The proportion with which a feature is observed in the set $\hat{G}$ is an approximation to its posterior probability within G:

$$\Pr(f|D) \approx \frac{\sum_{G \in \hat{G}} \Pr(G|D) f(G)}{\sum_{G \in \hat{G}} \Pr(G|D)} \tag{3.17}$$

As in Madigan and York [144], a Metropolis Hastings sampling algorithm is applied to generate the sample $\hat{G}$.

3.5 Metropolis Hastings Algorithm

Metropolis Hastings (MH) sampling is a Markov Chain Monte Carlo (MCMC) method that obtains samples from some probability distribution [150]. Here, samples are taken from the posterior distribution of Bayesian network topologies given data. There are two main reasons for the use of a Metropolis Hastings sampler in this application. Firstly, the sampler requires the computation of only a rela-

tive probability measure, a large advantage in this application, as the evaluation of the normalisation constant is only possible exhaustively. Secondly, MH sampling prefers the more important networks, that make a high contribution to the total probability integral of all possible networks. In most cases, networks that make a significant contribution are likely to be rare, which makes approaches such as random sampling inappropriate.

The chain proceeds through a finite number of steps, the current step is designated t. The state of the chain at $t + 1$ depends only on the previous state H^t. A move is proposed from the DAG H^t, generating a new DAG H'. The probability that the new DAG is accepted is the ratio of the DAG evidence scores (3.14), weighted the ratio of proposal ($Q(H^t; H')$) and reversal ($Q(H'; H^t)$) densities to and from the new DAG. As the probabilites are always relative, the normalising constant of the topology prior discussed above has no influence. This requirement is known as *detailed balance*, and is necessary to avoid bias owing to non-identical proposal and reversal probabilities. A proposed DAG is accepted as $H^t + 1$ if υ drawn from $U(0, 1)$ satisfies:

$$\upsilon < min\left\{\frac{\Pr(H')Q(H^t; H')}{\Pr(H^t)Q(H'; H^t)}, 1\right\} \tag{3.18}$$

If the transition probability is symmetric equation 3.18 simplifies to:

$$\upsilon < min\left\{\frac{\Pr(H')}{\Pr(H^t)}, 1\right\} \tag{3.19}$$

To ensure samples are drawn from the true topology distribution, it is important that the current state of the Markov Chain is irrelevant to its starting point. When this point is reached it the Markov Chain is said to be in equilibrium. Chains with different starting points should become indistinguishable and converge upon the most probabilistically dense regions of the space. A burn-in period is allowed for this to occur, however convergence may be problematic in some cases. Methods for checking and encouraging the mixing and convergence of MCMC chains are discussed in the next section.

Chapter 4

Software Development and Implementation

This chapter discusses the specifics of the implementation of the methodology described in the previous chapter. Much of the the chapter outlines the difficulties associated with Metropolis Hastings sampling over network structures, and describes how these issues are countered. The chapter also specifies the precise nature of moves between network structures. The sampling methodology used in this thesis is computationally intensive; the final section of the chapter discusses the heuristics of the approach, including approximations and limitations in order for the implementation to be tractable. The implementation is written as a C# program, which is made fully available, see appendix C.1.

4.1 Considerations of Mixing and Convergence over Network Structures

The Metropolis Hastings approach described in the previous chapter samples network topologies (H) from the posterior distribution of network topologies (G), the resulting sample of topologies is denoted $\hat{G}$ and represents an approximation to G. In order to sample from the true posterior distribution, the sampler must be able to explore the space thoroughly with free movement between topologies. If the algorithm is unable to successfully propose a move from a topology or set of topologies, this will lead to erroneous results and poor models. Successful transition between topologies in the space G is termed *mixing*. The posterior of network topologies G can be described as a *probability landscape*- in this instance a complex and multidimensional space. In practice, adequate mixing over network topology space is difficult to achieve, and is a recognised issue within the field [124, 144].

As the number of data points increases, the contribution of the evidence grows larger in relation to the prior. The probability landscape becomes more *peaked*; the absolute difference in evidence scores between topologies becomes greater. Consequently, moving to a Directed Acyclic Graph (DAG) with lower probability becomes relatively more difficult when the volume of data is greater. In order to move from a region of relatively low probability to a peak, it may be necessary to traverse a 'valley' by making several consecutive moves to less probable DAGs. In a model with a large amount of data this is more unlikely as the valleys are relatively deeper and the peaks relatively higher, thereby decreasing the probability of such moves and increasing the tendency of the Markov Chain to become stuck in local optima.

This issue is particularly pertinent when dealing with epidemiological data

which typically contain large numbers of observations. Many of the applications of Bayesian graph discovery in domains such as analysing gene expression data [107] involve datasets with large numbers of variables but few observations. Applying these techniques to epidemiological data requires careful consideration of potential mixing problems.

Not only is the probability landscape vast (eq. 3.16) and peaked, it is also highly complex. The probability landscape of network topologies may have numerous distinct peaks (local optima), even when the number of nodes in the network is relatively small. The space of network topologies is discrete which may be problematic for MCMC techniques. As there are no intermediate steps between DAGs, transitions between states are stochastic. A seemingly minor change such as the addition or deletion of a single arc may dramatically alter the evidence score of a topology. Topologies that differ by a single arc are not necessarily more likely to have a similar evidence score than two very distinct networks. Consequently, the concept of adjacency between DAG topologies is difficult. Instead, adjacent networks are defined by the networks reachable using a given move set, this composes the *neighbourhood* of DAGs.

In order for the MCMC sampler to obtain a representative sample over the posterior of network topologies it must be able to mix freely between regions of local optima and for multiple chains to converge reliably to regions of high probability. A number of approaches have been suggested to improve mixing and convergence of MCMC over network topologies, and are discussed in the next section.

4.2 Improving Mixing over Network Topology Space

4.2.1 Novel Moves

The operations available to move between network topologies define the *neighbourhood* of a DAG; the set of DAGs that can be reached in a single transition. Although the space G is invariant, the probability landscape traversed is determined by the available moves, as the moves define which networks are reachable from a given topology. Early implementations of MCMC over the space of network topologies (e.g. [144]) relied on very simple moves between topologies; the addition, deletion and reversal of single edges. These simple moves are referred to as *Classical* or *Structural* MCMC by Grzegorczyk and Husmeier [143]. In theory, these moves provide access to all possible DAGs- however they are likely to be

extremely inefficient at exploring the space of network topologies.

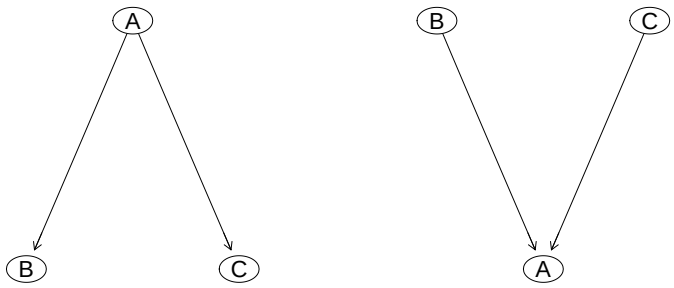


Figure 4.1: Diagram showing limitations of a simple scheme to move between network topologies

In order to illustrate this, consider a trivial network with 3 nodes: $\{A, B, C\}$. Assuming the optimal network has arcs $\{B, A\}$, $\{C, A\}$ (fig. 4.1(b)), it is impossible to reach this from the topology $\{(A, B), (B, C)\}$ (fig. 4.1(a)) using the *Classical* move set without visiting at least one intermediate state. If this state is of much lower probability than the original network the sampler is unlikely to traverse the space without a significant number of iterations. In more complex networks the number of intermediate states required may be substantially larger. The design of intelligent moves can at least partially circumvent this issue by allowing 'shortcuts' between high probability regions of topology space.

One such move is the recently developed Grzegorczyk-Husmeier edge reversal (REV) move [143], which is implemented in this thesis. An arc is selected at random, reversed, and all incident arcs deleted and parentsets are resampled. Section 4.4.2 discusses the implementation and effect of this move. By allowing moves between high scoring networks, the REV move extends the neighbourhood of DAGs, resulting in a more easily traversable probability landscape.

Castelo and Kocka [151] proposed a modified MCMC approach reliant on equivalence classes of network structures. An equivalence class is a set of networks that although different, fundamentally specify the same conditional dependencies [152]. The proposal mechanism efficiently selects a DAG from the inclusion boundary which is defined as the set of all DAGs that can be reached by

a classical move (add, delete or reverse an arc) from any member of the current equivalence class. This results in a more easily traversed probability landscape due to the ability to make slightly longer range moves by 'leapfrogging' other members of the equivalence class. Grzegorczyk and Husmeier suggest this approach assists in travelling along probability ridges, rather than traversing valleys [143].

The set of moves available are a very important determinant of mixing and convergence, the design of novel moves allows transition to high scoring topologies while traversing fewer probability valleys, thus improving mixing.

4.2.2 Partitioning the Space of Network Topologies

The complexity of the space G can be reduced by partitioning the space into smaller sections. Friedman and Koller [124] developed a two stage approach using node orders. A node order is a sequence of nodes such that a node can only have preceding nodes as parents. Given a specified node order, the space of potential DAGs is greatly reduced. The key component of this approach is the ability to analytically evaluate the probability of a given node order. Firstly, MCMC is applied over the space of node orders; a conventional MCMC algorithm is then applied over DAG space subject to the defined node order.

This approach yields drastically improved mixing and convergence due to the less peaked space within a node order. However, several authors have noted that due to difficulties with specifying a prior over node orders, the method does not accurately represent the posterior distribution of topologies; Ellis and Wong provide a succinct explanation in [153]. This bias can be overcome with the use of specially designed algorithms, as developed by Kovisto and Sood [154, 155], and Eaton and Murphy [156]. Grzegorczyk and Husmeier argue that for large networks ($\geq$20 nodes) these approaches become computationally too expensive, and are not appropriate for biological applications and in other domains that may use large networks. In addition, this approach does not allow the inclusion of explicit prior knowledge via the inclusion or exclusion of specific arcs *a priori* which may be desirable in some cases.

4.2.3 Programmatic Methods

Some general methods exist for improving the dynamic properties of MCMC analyses. One such approach is *Parallel Tempering*, where multiple chains are run at different temperatures T. As outlined in section 3.5, Metropolis Hastings involves

a stochastic process that rejects or accepts updates with probability based upon the evidence ratio of the proposed state to the current state, weighted by the proposal and reversal probabilities. The temperature T is a multiplicative factor that influences the probability of proposal acceptance. In a standard Metropolis Hastings algorithm $T = 1$, by increasing the value of T the probability of acceptance is increased, this makes probability valleys far less deep. Parallel Tempering periodically exchanges states between runs of different temperatures. However, as samplers using different values of T are effectively sampling from different posterior distributions, maintaining the detailed balance of the system is problematic. Both Linderman [157] and Asadi *et al.* [158] have applied Parallel Tempering within the space of node orders.

Stochastic Application Monte Carlo (SAMC) is a variant Metropolis Hastings algorithm described by Liang *et al.* [159, 160]. This is a dynamic algorithm that is designed to artificially 'jump' to a distinct region of topology space when it has become stuck. However, it requires the user to define sub regions of DAG space that are likely to contain distinct probability peaks. This may not be feasible for large networks.

4.2.4 Approximation of the Space of Network Topologies

The plethora of available technologies may not be able to overcome the mixing and convergence challenges presented by a highly peaked and complex probability landscape. The datasets used in this thesis are large- substantially larger than those used for the testing of the REV move and order MCMC [124, 143], and other typical applications of the technology (e.g. high-throughput microarray data [157]). Given a highly peaked landscape, the sample may be restricted to networks in the neighbourhood of the global optimum by seeding the Metropolis Hastings sampler with the optimal DAG. This approach is reasonable subject to two main assumptions:

- Firstly that the distribution of network topologies is highly peaked such that the region around the global optimum is sufficiently probabilistically dense as to form a satisfactory approximation to the entire space G.
- Secondly that we can be confident that the true global optimum is identified successfully.

The most difficult condition to meet is the first, which is problematic if there are two distinct regions that possess comparable probability integrals. Further, it will

be difficult to determine whether the sampler leaves the area of high probability, and becomes trapped in a distinct region. This is an approximation to sampling from the true posterior, and is not an ideal solution.

As the number of data points approaches infinity, the probability landscape becomes a Dirac delta distribution, peaked at a single network topology. In such instances when adequate mixing and convergence is unobtainable it may be more informative to obtain the optimal DAG, this approach becomes more reasonable with increasing data.

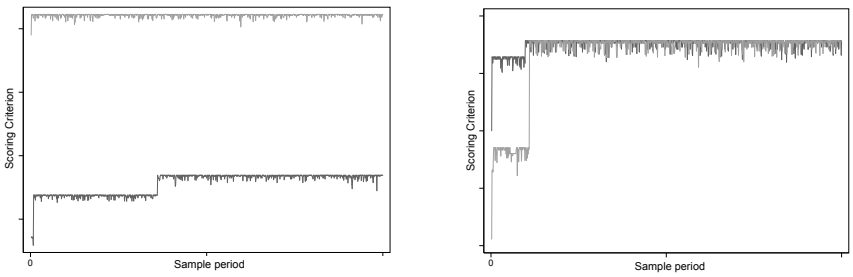
4.3 Monitoring Mixing and Convergence

As previously noted, mixing and successful convergence of MCMC chains is essential to derive an accurate representation of the distribution of interest. It is also known that mixing over topology space is problematic; prone to becoming stuck in regions of low probability. It is important therefore to monitor mixing and convergence to ensure sampling is taking place from the correct distribution.

Metropolis Hasting sampling is an importance sampling approach, the most important (*i.e.* probabilistically dense) regions are traversed most thoroughly. Multiple chains should converge on the highest scoring regions of topology space. In practice, mixing and convergence is evaluated by the initiation of chains from different starting points and monitoring their convergence. Quantitative measures of convergence are available, these provide a measure of the overlap between sampling chains [161]. However such measures are subject to interpretation, and are ultimately judged according to the investigator's discretion [162].

Visualisations are often used in conjunction with quantitative measures. As well as being more intuitively understood, visualisations allow more opportunity to understand reasons underlying convergence problems. The simplest visualisation is the plotting of the scoring criterion of multiple chains over the sampling period. In this case the evidence criterion of the network topology is used. Where convergence is achieved, chains will reach the same areas of the space, and will display similar evidence scores (e.g. fig. 4.2(b)). However, if the chains fail to overlap (fig. 4.2(a)), this indicates a mixing problem, where one chain has become stuck. By examining when chains tend to converge, this provides an indication of how long the sampler takes to 'burn-in' *i.e.* when the state of the sampler becomes independent of its starting state.

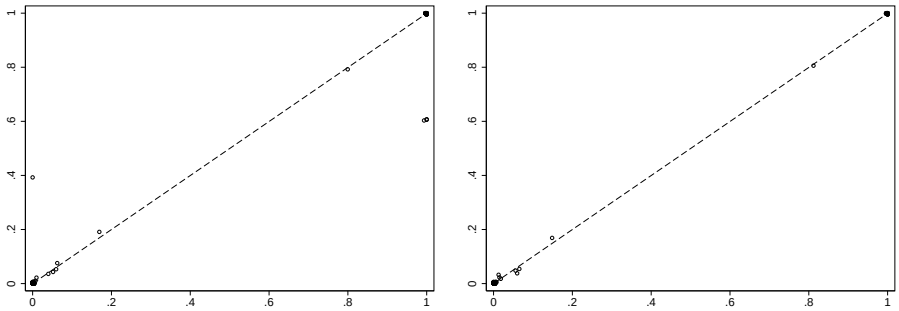
Another simple yet informative visualisation are pairwise scatter plots, which



(a) Unsuccessful Convergence (b) Successful Convergence

Figure 4.2: Evidence traces of two Markov chains as examples of successful and unsuccessful convergence

plot features of two MCMC chains (following burn-in) against one another. For MCMC over network topologies the plotted parameters are the edge relation features (ERFs) described in the previous chapter. The ERF of each pairwise combination of nodes represents a parameter. When the number of parameters is large, this visualisation becomes less attractive. If the chains have successfully converged, agreement between chains will be high, resulting in very high positive correlation on a scatter plot. It is not appropriate to calculate a correlation coefficient as a single outlier indicates mixing has not been successful. Examples of scatter plots are shown in figure 4.3.



(a) Unsuccessful Convergence (b) Successful Convergence

Figure 4.3: Scatter plots of features of two Markov chains as examples of successful and unsuccessful convergence

The described techniques monitor *convergence*; crucially they do not necessar-

ily imply that the sampler has successfully reached the most probabilistically dense region. In order to verify that a sampler has successfully reached this region, optimisation techniques can be applied to find the peak of the distribution. Although this doesn't guarantee finding the optimal region, it can indicate that there are better regions in the event that the chains have converged upon a shared but inferior space.

Throughout this thesis the two visualisation techniques of evidence scoring criterion traces and agreement scatter plots are used to *qualitatively* monitor mixing and convergence.

4.4 Implementation of Moves between Network Topologies

4.4.1 Move Library

This section describes the moves implemented by the Metropolis Hastings sampler applied in this thesis. This move set defines the process by which a new DAG (state) is generated. The choice of move set is extremely important, as the available moves determine the effective probability landscape of the space (4.2.1). To fulfill the requirements of detailed balance, Metropolis Hastings sampling requires the knowledge of the proposal and reversal probabilities for each move. Consequently, it is important that each move in the set is mutually exclusive; that a transition from a state t to $t+1$ is impossible to reverse unless performing the reciprocal move. The moves implemented are listed below:

- Add Arc. An arc is selected with uniform probability from a list of permitted arcs (*i.e* those that form a valid DAG).

- Delete Arc. An existing arc is selected with uniform probability and deleted.

- Grzegorczyk-Husmeier REV move (described in [143]). The REV move selects an edge $X_i \rightarrow X_j$ at random, reverses it while deleting all edges incident into X_i and X_j. New parent sets are then sampled for X_i and X_j. Section 4.4.2 provides more detail.

- Switch Arc. The Add and Delete functions above are used to add an arc, then immediately delete one. This order ensures no arc is reversed.

$\tilde{H} \leftarrow H$	$N^{\dagger}_{\tilde{H}}$	$M^{(\{\tilde{H} \leftarrow H\})}$
Add/Delete Arc	$N^{\dagger}_{H} \pm 1$	0
REV move	varies	1
MR move	$N^{\dagger}_{H}$	2+
Switch Arc	$N^{\dagger}_{H}$	0

Table 4.1: Table showing how implemented moves are mutually exclusive by comparing features of resulting networks

- Multiple Reversal (MR) move. This is a novel move; a node with at least 2 adjacent nodes is selected and all associated arcs are reversed, subject to acyclicity constraints. The reversal of at least 2 nodes again ensures no possible overlap with the REV move.

We can guarantee that the moves used in this thesis are mutually exclusive. Consider $N^{\dagger}_{H}$ as the number of arcs in topology H, and $M^{(\{\tilde{H} \leftarrow H\})}$ as the number of arcs reversed in the transition $\tilde{H} \leftarrow H$.

Table 4.1 shows the different properties of the DAG $\tilde{H}$ following the transition $\tilde{H} \leftarrow H$ following each move implemented. Using the REV move as an example, to generate the new DAG $\tilde{H}$, *exactly* one arc in H has been reversed. From table 4.1, only the REV move is able to generate a new DAG with one arc reversed from the original DAG. Consequently the only move that can make the transition $\tilde{H} \rightarrow H$ is REV. This logic can be extended to show all moves are mutually exclusive.

4.4.2 Implementation of the Grzegorczyk-Husmeier Reversal Move

The Grzegorczyk-Husmeier move is a recently developed edge reversal move that has been shown to improve mixing in MCMC processes [143]. The move generates a new DAG $\tilde{H}$ from the existing topology H, and proceeds as follows:

(1) From the graph H an arc $X_i \rightarrow X_j$ is selected at random. All arcs into nodes X_i and X_j are deleted, providing the DAG: $H_{\odot} := H^{\{X_i, X_j\} \leftarrow \emptyset}$.

(2) A new parentset $\tilde{\pi}_i$ for X_i is sampled from the set of all permitted parentsets, the probability of selecting a given parentset is weighted by the local score (*i.e.* evidence score of topology of node and parents only (eq. 3.12)) of each potential parentset given observed data. Potential parentsets are determined by two criteria; they must form a valid DAG, and contain the node X_i. The

probability of drawing the parentset $\tilde{\pi_i}$ for node X_i given data D is given by:

$$Q(\tilde{\pi_i}|M_\odot, X_i) = \frac{\exp \psi[X_i, \tilde{\pi_i}|D]}{\sum_{m=1}^{n} \exp(\psi[X_i, \tilde{\pi_m}|D])}. \tag{4.1}$$

where $\psi[X_n, \tilde{\pi_n}]$ represents the evidence score of X_n given $\tilde{\pi_n}$. This yields the DAG: $H_\oplus$.

(3) Finally, a new parentset $\tilde{\pi_j}$ is similarly sampled for the X_j. Here the set of potential parentsets for X_j is limited only by the acyclicity constraint (*i.e.* formation of a valid DAG, with no cycles). Hence an empty parentset is possible.

The proposal probability is therefore:

$$Q(H; \tilde{H}) = \frac{1}{N_H^\dagger} Q(\tilde{\pi_i}|H_\odot, X_i) Q(\tilde{\pi_j}|H_\oplus, X_j). \tag{4.2}$$

where $N_H^\dagger$ is the number of arcs in topology H.
The reversal probability $Q(\tilde{H}; H)$ is similarly calculated in stages. In the reciprocal move, the roles of X_i and X_j are reversed; the move is performed on the arc $X_j \rightarrow X_i$:

(1) All arcs incident to X_j and X_i are deleted, this returns the graph: $H_\odot := H^{\{X_j, X_i\} \leftarrow \emptyset}$.

(2) The probability that the parentset π_j of X_j present in H is sampled is given by:

$$Q(\pi_j|H_\odot, X_j) = \frac{\exp \psi[X_j, \pi_j|D]}{\sum_{m=1}^{n} \exp(\psi[X_i, \tilde{\pi_m}|D])}. \tag{4.3}$$

The set of potential parentsets is restricted by the inclusion of X_i in each π_m, and the usual constraints of acyclicity.

(3) The probability that the parentset π_i of X_i present in H is sampled is calculated in the above manner, except the set of potential parentsets is restricted only by acyclicity.

The reversal probability $\tilde{H} \rightarrow H$ is therefore:

$$Q(\tilde{H}; H) = \frac{1}{N_{\tilde{H}}^\dagger} Q(\pi_j|H_\odot, X_j) Q(\pi_i|H_\oplus, X_i). \tag{4.4}$$

As noted by Grzegorczyk and Husmeier [143], the unrestricted REV move is computationally intensive. The number of potential parentsets that must be evaluated for each move (λ) rises exponentially with the number of eligible parents (μ):

$$\lambda = 2^{\mu(X_i|H_\odot)} + 2^{\mu(X_j|H_\oplus)}. \tag{4.5}$$

A similar number must be computed to calculate the reversal probability:

$$\lambda = 2^{\mu(X_j|H_\odot)} + 2^{\mu(X_i|H_\oplus)}. \tag{4.6}$$

Further, parentsets with more members have a large number of input levels, which slows calculation of the local score and can cause memory issues. Heuristic methods to increase tractability are discussed in sections 4.6.

Experimental validation of the REV move

Although the REV move has been previously shown to improve mixing [143], its use is justified in this different context by analysing the mixing properties of Metropolis Hastings sampling over the topology space of a dataset used in chapter 7. This dataset contained 15 discrete variables and 3,281 observations, far in excess of the datasets used for previous evaluation [143]. The REV move was developed primarily for use in the field of gene expression analysis, where datasets contain continuous variables and are much broader and smaller, *i.e.* more variables and fewer observations. Consequently, application of the REV move to much more peaked discrete datasets has not been evaluated.

Mixing and convergence is compared between a classical MCMC scheme, (*i.e* where addition/deletion of edges and a simple reversal move only are implemented), and a scheme involving the REV move (replacing the simple reversal move). The *Classical* uses the following move frequencies; Add arc- 0.4, Remove arc- 0.4, Switch arc- 0.1, Simple reversal- 0.1. The REV scheme replaces the Simple reversal move with the REV move described in the previous section. For clarity, the Simple reversal move determines which arcs can be reversed to form a valid DAG, then selects one to reverse with uniform probability. Despite differences in computational speed between the REV and the simple reversal moves, no allowance was made for this in terms of the number of iterations allowed to

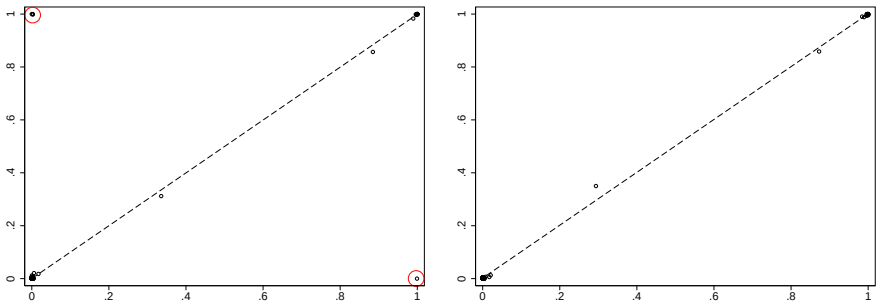
achieve convergence. This was because the time taken in each case was not felt to be a limiting factor of the analyses. Each run of both schemes was conducted for a burn in period of 5×10^5 iterations, before a sampling period of a further 5×10^5 iterations. Samples were taken every 1000 iterations during this period, providing a total sample of 500 DAGs. 4 independent runs of the sampler were performed; 2 from an empty DAG with no edges, and 2 from the optimal (informed) DAG generated using simulated annealing (see section 4.7).

Each run of the Metropolis Hastings sampler provides a sample of network topologies $H_1, ..., H_n$. As described in section 3.4, Bayesian Model Averaging is used to identify common features of the sampled topologies. Here we examine edge relation features, an estimate of the posterior probability of an edge over the space of network topologies G. When estimating the edge relation features of a set of topologies the direction of arcs is disregarded. The posterior probability of an edge between two nodes is estimated by counting the proportion of sampled DAGs in which the edge is present (eq. 3.17).

Each sample of network topologies provides a set of estimated probabilities for each pairwise combination of nodes. If mixing and convergence of the Metropolis Hastings chains is good, the resulting feature estimates of independent runs will be well correlated. Estimates of edge relation features between runs are plotted against each other in figure 4.4.

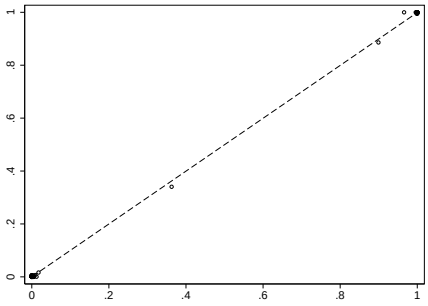
Trace plots of the evidence score can be seen in Appendix C.2. Scatter plots comparing the results of MH chains are seen in figure 4.4. The plots of the Classical scheme show very poor correlation 4.4(a), unless initialised from the informed graph 4.4(c). The lack of convergence is also evident in the evidence score traces in Appendix C.1(a). In contrast, the REV move shows a high level of correlation between the two runs from both an empty network and the seeded optimal network (4.4(b) and 4.4(d)). In figure 4.4(e) an empty initialisation and the informed initialisations of the Classical scheme are compared; the empty runs are not combined as they failed to converge; initialisation of the scheme is important. This is in contrast to the REV scheme (figure 4.4(f)), where we see that the initialisation does not influence the results obtained. The Classical sampler is capable of returning results consistent with the REV sampler only when seeded with an informed DAG, and when the conditions outlined in 4.2.4 are met.

It is notable how bimodal the distribution of posterior probability estimates is, with the vast majority of points being close to 0 or 1. This is also described by Grzegorczyk and Husmeier [143] and is due to reduced inference uncertainty

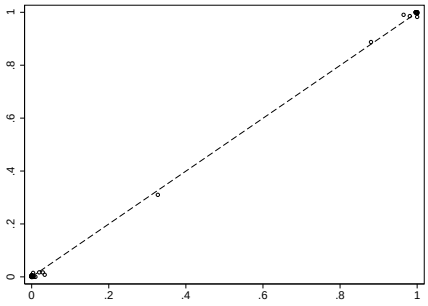


(a) Classical (empty-empty)

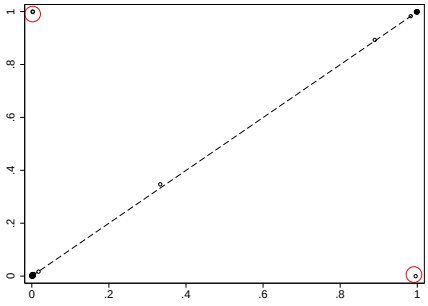
(b) REV (empty-empty)



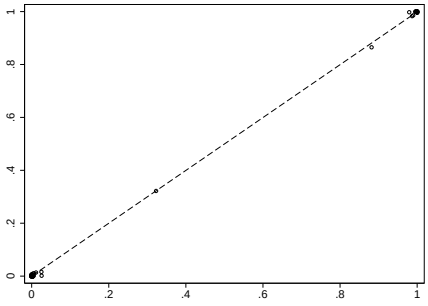
(c) Classical (informed-informed)



(d) REV (informed-informed)



(e) Classical (empty-informed)



(f) REV (empty-informed)

Figure 4.4: Scatter plots of edge relation features to compare convergence of Markov chains between schemes (classical vs REV)

with increasing dataset size. The impact of the Grzegorczyk-Husmeier REV move is impressive; as noted previously the dataset applied here far exceeds those investigated in the original paper [143]. The implementation of this move helps to overcome significant barriers to the research described in this thesis.

4.4.3 Implementation of the Multiple Reversal Move

In addition to the REV move, a further novel move is also implemented to promote better mixing. Following analysis of some early runs of the Metropolis Hastings sampler, it was noticed that chains were not converging to the same evidence scores, but were displaying similar network structures, with minor differences. These differences resulted in significant disparities between evidence scores. Upon closer inspection, in many cases the difference was a 'V' structure common in one set of DAGs being inverted in the other, as illustrated in figure 4.5.

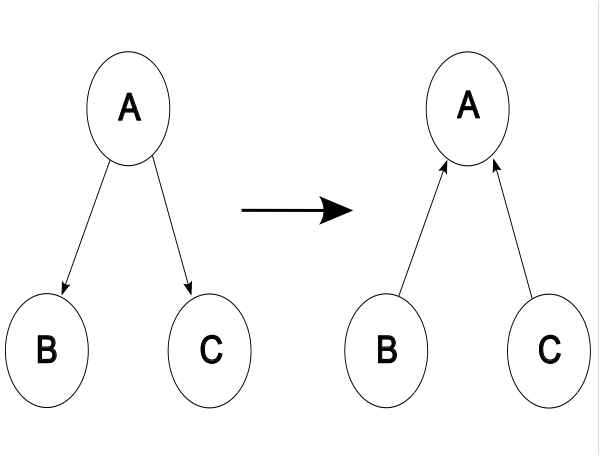


Figure 4.5: An example of a transition not directly possible using classical or REV moves

Neither the Classical nor REV moves can move between these two DAGs directly. The *Multiple Reversal* (MR) move was designed to allow a direct transition aiding mixing between these regions. Briefly, an eligible node is selected and all associated arcs are reversed.

The MR move from a DAG H to the new DAG $\tilde{H}$ proceeds as follows:

(1) The set of eligible nodes is identified and denoted V_H. Eligible nodes must a) have at least 2 adjacent nodes (*i.e.* parents or children), and b) the reversal

of associated arcs must result in a valid DAG. If the size of this set is zero, a new move is selected.

(2) A node X is selected from the set V_H with uniform probability. The proposal probability is simply $(N_{V_H})^{-1}$. All arcs incident to and from X are reversed- the resulting DAG is $\tilde{H}$.

(3) To calculate the reversal probability, step 1 is repeated for the DAG $\tilde{H}$; $(N_{V_{\tilde{H}}})^{-1}$ provides the reversal probability. We can be sure that the set $V_{\tilde{H}}$ has at least one member, as the application of the MR move to the node X results in the original (and valid) DAG H.

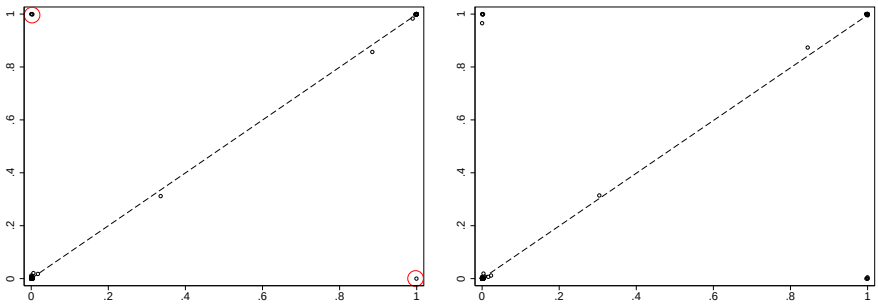
Experimental validation of the multiple reversal move

Like the REV move the MR move has the effect of enabling a 'shortcut' between distinct regions of topology space. Although the transition described is possible over two steps using the classical move set, a peaked probability distribution will make this transition far harder, assuming the intermediate state is low scoring.

The experimental validation performed here compares the *Classical* move set described above, with a move set including the MR move. The MR move is added to the list of possible moves; Add arc- 0.4, Remove arc- 0.4, Switch arc- 0.05, Simple reversal- 0.05, multiple reversal- 0.1. Neither scheme included the REV move.

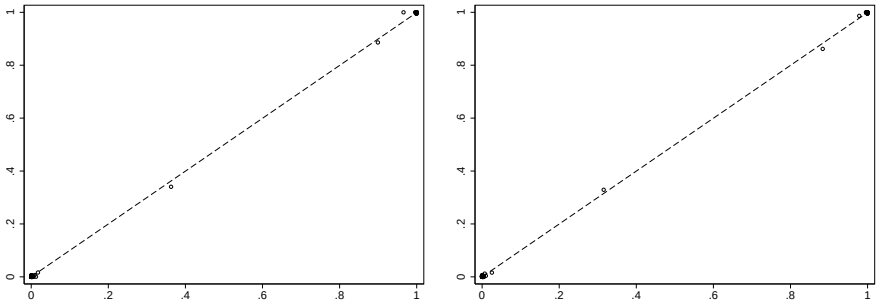
The above methodology was applied; each run of both schemes was performed for a burn in period of 5×10^5 iterations, before a sampling period of a further 5×10^5 iterations. Samples were taken every 1000 iterations during this period, providing a total sample of 500 DAGs. 4 independent runs of the sampler were performed; 2 from an empty DAG with no edges, and 2 from the optimal (informed) DAG generated using simulated annealing (see section 4.7).

Evidence traces can be seen in the appendix (C.3). Neither scheme successfully converged on the highest scoring region from an empty DAG (fig. 4.6). Although it was not possible to validate the MR move experimentally here, it may have value in specific instances. It is nonetheless included in the implemented move library, as in theory it enables movement between DAGs not permitted by the classical and REV moves, and can be shown not to disrupt detailed balance. The usefulness of each move will depend on the nature of the probability landscape, it is impossible to evaluate the utility of a move in all contexts.



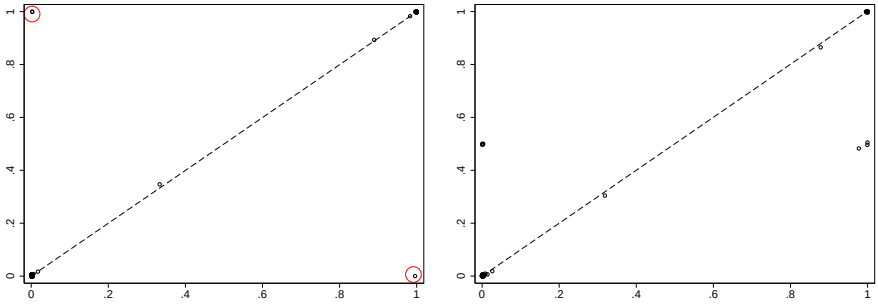
(a) Classical (empty-empty)

(b) MR (empty-empty)



(c) Classical (informed-informed)

(d) MR (informed-informed)



(e) Classical (empty-informed)

(f) MR (empty-informed)

Figure 4.6: Scatter plots of edge relation features to compare convergence of Markov chains between schemes (classical vs MR)

4.5 Validation of DAGs

The operations available to move between topologies described in the previous sections require a method of ascertaining whether or not a particular network structure is valid. Often, it is necessary to perform this operation on a large number of structures, for example when determining which arcs currently not present form a valid DAG when added. Consequently it is important that the operation can be performed quickly and efficiently. Bayesian networks take the form of Directed Acyclic Graphs (DAGs), hence all arcs are directed and the graph must contain no cycles. It is permissible for some nodes or sets of nodes not to be connected to the main graph.

Networks are represented within the program as adjacency matrices, a square matrix where the index [i, j] has values 0 or 1, representing the absence or presence of an arc from node i to j. Given a network, the validity-checking function moves through each node in the topology. When it encounters a node with no children, this node is deleted, and all incident arcs removed. The algorithm then moves on to the next existing node, and repeats the process. If the process completes a loop of all remaining nodes without a deletion, the structure contains a cycle and is not a valid DAG. Otherwise, deletion of all nodes shows that the structure is acyclic.

4.6 Tractability of Metropolis Hastings Sampling

4.6.1 Efficient Updating of Network Evidence

The scoring criterion used throughout this thesis to evaluate networks is shown in eq. 3.12. It can be thought of as the product of each marginal likelihood (eq. 3.11) over each node and input level given data, multiplied by a prior on the network topology (eq. 3.13). Given a network structure, the marginal likelihood score of each node-input combination can be computed independently.

For each marginal likelihood score calculated at node i and input level j, it is necessary to generate an array of integers from data that represent the counts at each outcome level k. This process takes substantially more processing time than the evaluation of the marginal likelihood function in 3.12. In a highly connected network, where several nodes have multiple parents, the number of input levels may become large, increasing the computational time to generate counts.

As eq. 3.12 depends upon these counts, the marginal likelihood associated with each ij combination only varies with the parentset of the node i. Consequently, to

improve efficiency, a memory of the marginal likelihood contribution each node is maintained. The contribution of each node is the product of the marginal likelihood scores at each input level. Whenever the parentset of a node is updated, the marginal likelihood contribution of that node is re-evaluated. This method significantly improves the efficiency of topology scoring.

Validation of Node Likelihood Contribution Updating

Computational time savings of the above implementation are evaluated here. Precise timings are likely to be dependent on the connectivity of networks visited, as likelihood calculation of nodes with more input levels is more time consuming.

Two Metropolis Hastings samplers were run, each initialised with the same starting network, over the dataset used in chapter 7 and previously used in section 4.4.2 of this chapter. The dataset contains 15 discrete variables and 3,281 observations. The first sampler uses the *original* method, where no memory of the likelihood contribution of nodes is maintained. The second *memory* method maintains such a list. Seeds were set so that both samplers explored exactly the same space over the course of their MCMC runs. The cumulative time taken to perform each

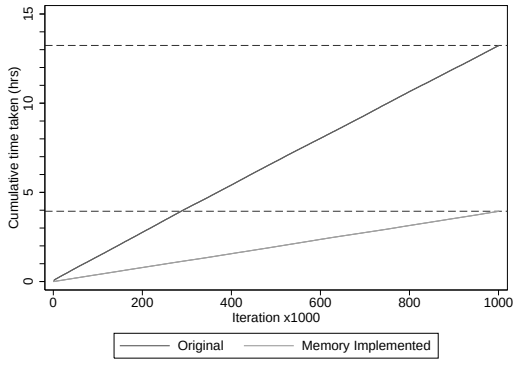


Figure 4.7: Improved efficiency of evidence calculation using contribution updating compared to re-evaluation of whole network

sampler is displayed in figure 4.7. The implementation of this method has the effect of reducing the time taken to run the MCMC sampler by approximately 70%. The original sampler took 13.23 hours compared to 3.94 for the more efficient method.

4.6.2 Caching of Parentsets for the Grzegorczyk-Husmeier Move

The evaluation of parentset local scores necessary for the REV move is computationally expensive. The time taken to evaluate a local score of a single parentset varies linearly with the multiplicative factor of the number of observations and the number of input levels. Subsequent calculation of local scores of previously evaluated parentsets represents a great deal of redundancy. This extremely inefficient in an MCMC approach requiring $\sim 1 \times 10^6$ iterations. As $\psi[X_z, \pi_z|D]$ is invariant given D, the local score associated with each parentset can be stored and recalled when required. Without this functionality MCMC runs become intractable for most reasonably complex networks.

In the C# implementation of the Metropolis Hastings sampler, local scores of potential parentsets for each node are cached in a Sorted Dictionary object, indexed by a bit-key. The bit-key ϕ is generated by ranking all other nodes and indexing them 0 to $n-1$:

$$\begin{aligned} \phi(\pi_z) &= \sum_{i=0}^{n-1} (2^i)^{\mathbb{I}_{\pi_\mathbf{z}}(i)} \\ \mathbb{I}_{\pi_\mathbf{z}}(i) &= \begin{cases} 1 & \text{if} \quad i \in \pi_\mathbf{z} \\ 0 & \text{if} \quad \text{otherwise.} \end{cases} \end{aligned} \tag{4.7}$$

The authors of the REV move suggest prior evaluation of all parentsets [143] as a means of reducing overall processing time. The approach taken here is preferable as parentsets are evaluated as required, resulting in less redundancy. Further, the storage of all possible parentsets across all nodes may encounter memory issues. I believe the small overhead associated with checking whether a parentset has been evaluated is a worthwhile price for this reduced redundancy.

All bit-key indexed parentset local scores are written to file at the end of a session, ready to be loaded into memory for the next. This maintains a continuously updated library of parentset local scores. This is summarised in pseudocode:

```
Begin session
Determine dataset characteristics (filename,#observations)
Check if a file of current dataset exists.
 True: Load into memory.
 False: Initialise empty array of n sorted dictionaries
loop{
 Generate potential parentset, determine bit-key;
  check if already evaluated.
```

```
    True: look up key and use local score.
    False: evaluate.
           Add bit-key/score to dictionary
End session.
Write array of Sorted Dictionaries to file.
```

Validation of parentset caching

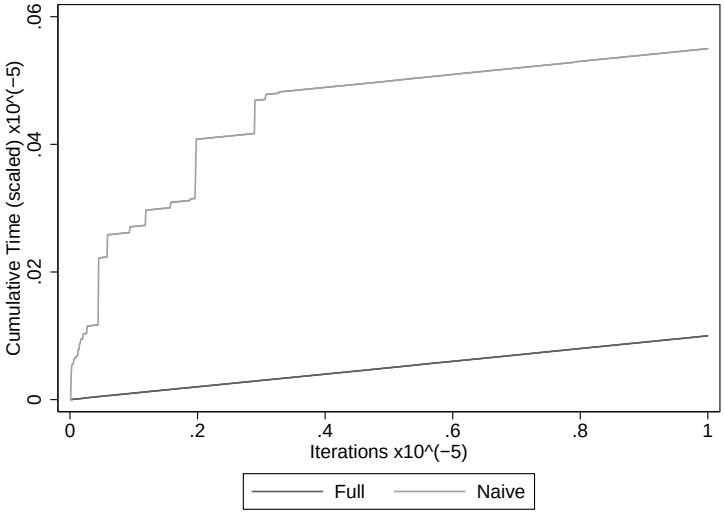
Here, the necessity of parentset caching is shown under experimental conditions. This validation was performed on a benchmark dataset rather than a dataset from this thesis. This was because parentset caching was investigated in the exploratory stages of this analysis, before datasets were finalised. The UCI mushroom dataset was used [163] (23 categorical variables, 8,432 observations attenuated to a random set of 2,000 for convenience), the MH sampling algorithm was applied over the space of possible network topologies.

Two different versions of the algorithm were compared:

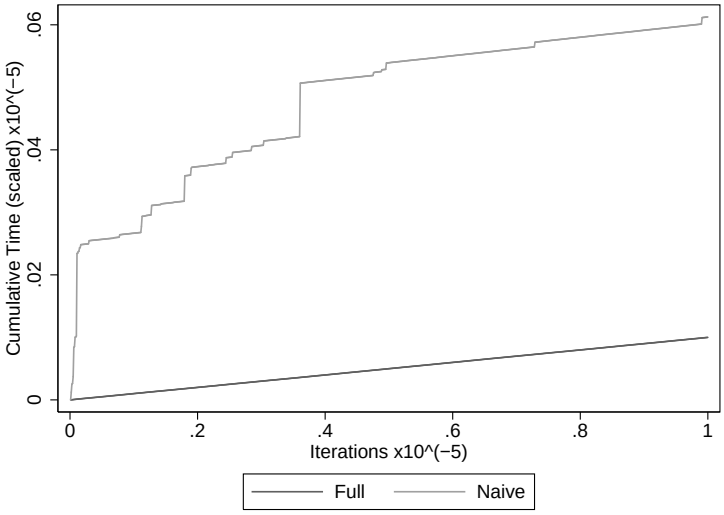
- **Naïve.** Caching and recall enabled, the process began with an empty parentset record.
- **Full.** The process began with a record of all parentsets found using the *Naïve* algorithm.

Each version of the algorithm was tested over 2 different starting networks A & B. By setting a seed in the random number generator it is ensured that the runs A & B are exactly equivalent for each version of the algorithm, *i.e.* they share the same starting networks and evaluate the same parentsets throughout the run. It is worth noting that the *Full* setting will never have to evaluate any parentsets directly as the *Naïve* setting will have visited them previously. The results can be seen in figure 4.8. The graphs show the time taken by each version of the algorithm to complete 100,000 MCMC iterations. The y-axis is scaled to the mean completion of 100 iterations by the *Full* algorithm.

The graphs show the significant gain associated with caching of parentsets. The evaluation of new parentsets evidently represents a bottleneck in the execution of the program. The steep gradients represent the naïve algorithm entering a new region of DAG space where new parentsets are encountered; for example, a node may become an eligible parent following a move; this opens a tranche of new parentsets to be evaluated. Without parentset memory implemented the sampler's sluggishness makes it inviable to determine timings.



(a) Run A



(b) Run B

Figure 4.8: Computational gain of parentset caching in the REV move

4.6.3 Imposing a Reduced Candidate Set

Despite efficient caching, evaluation of parentsets remains computationally intensive. The large numbers of parentsets that must be evaluated is restrictive. As networks become larger increasing numbers of eligible parents mean the number of potential parentsets rises exponentially (eq. 4.5 & 4.6); an additional node in the set of eligible parents doubles the number of potential parentsets.

In order to restrain these numbers, the authors of the REV move suggest reducing the number of eligible parents for a given node by generating a set of nodes from which eligible parents must be drawn [143]. The members of this reduced candidate set are chosen by ranking the local score of the node in question (X_i) with each potential parent as the sole member of the parentset π_i. Thus eligible parents are drawn from a set of the most probable parents. Grzegorczyk and Husmeier do not provide results from the application of this approximation in their paper. However, the effects of implementing different sizes of reduced candidate sets on computational time are investigated here.

Such a restriction of DAG space represents a significant approximation. It is implicity assumed that all topologies that contain an arc between a pair of nodes $X_i \rightarrow X_j$ where X_i is not a member of the reduced candidate set of X_j are non-important. There is an obvious issue of sensitivity when this approximation is applied; if too restrictive, a significant proportion of the probability integral of the distribution over the space of topologies (G) may be ignored. Although not the primary intention, choosing parents from a reduced candidate set reduces the DAG space that the algorithm can explore, this may improve mixing in a similar manner to the order MCMC of Friedman and Koller [124].

The effects of imposing different sizes of reduced candidate sets on computational speed were investigated. Difficulties were encountered with the detailed balance requirement of Metropolis Hastings sampling where reduced candidate sets were not reciprocated, *i.e* where X_j was an eligible parent of X_i, but not vice versa. To illustrate this, let us suppose that in a DAG M, X_j is a parent of X_i. If the REV move is performed on the arc X_j to X_i, X_i becomes a parent of X_j and new parentsets π_i and π_j are sampled to form the new DAG $\tilde{M}$. Following a REV move performed on $\tilde{M}$ to $\tilde{\tilde{M}}$, when calculating the probability of reversal from $\tilde{\tilde{M}}$ to $\tilde{M}$ an error is generated, as X_i is not an eligible parent of X_j and such a move is impossible. Similar errors occur if X_i becomes a parent of X_j via any of the other moves. All eligible parents must therefore be reciprocated. Further, other moves must also

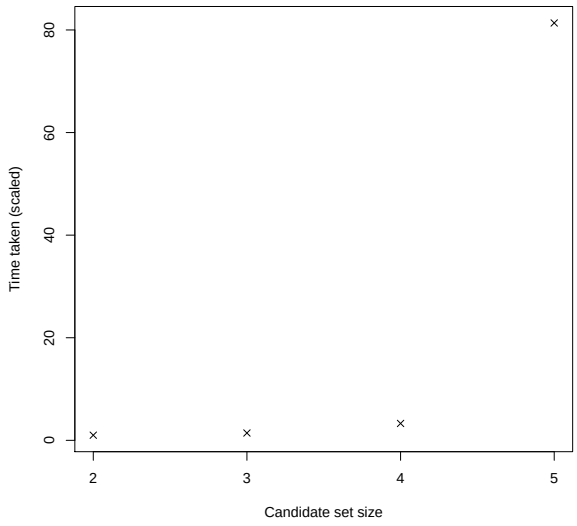


Figure 4.9: Influence of candidate set size on computation time

abide by the eligible parent restrictions. This reciprocation of edges means the size of the reduced candidate sets is not consistent between all nodes. Thus, the designation of a limit is not precise, with some influential nodes having many eligible parents.

Using the dataset described in the previous section (3,281 observations, 15 variables), several Metropolis Hastings samplers were implemented over the space of network topologies, applying a range of reduced candidate set sizes (2-5), beyond this the process was not tractable in reasonable time. The time taken to complete 10,000 iterations was recorded (move frequencies: 0.4 add, 0.4 remove, 0.1 switch, 0.1 REV). The results were scaled to the time taken to complete the simulation with a set size of 2 (41.3s), and are plotted in figure 4.9. This test was run on a 2.4GHz dual core machine with 2GB of RAM.

As shown in figure 4.9, the computational time required increases exponentially with the upper limit of the reduced parentset size. As equations 4.5 & 4.6, we would expect the computational time to roughly halve with every reduction in the set size. However, this exponential relationship will not continue throughout the sampling period, as the sampler will not continually explore new space; cached parentsets will be used. The uneven distribution of parenset sizes (discussed above), and the higher computational load of larger parentsets makes exact timings unpre-

dictable. The exponential relationship shows that some limit is required to make the method tractable, particularly as the network (and hence μ) becomes larger. The size of the reduced candidate set is a trade off between ignoring potentially important DAG space and computational speed. Without detailed investigation, the extent to which the results are compromised by imposing a parentset limit is unclear.

4.6.4 Restrictions on Node Cardinality

Grzegorczyk and Husmeier also discuss imposing a node cardinality limit, *i.e.* restricting the maximum number of parents a node can have. They argue that this is often justified in the case of gene expression data as genes are rarely regulated by more than a couple of other genes, but can regulate the expression of many others. The same assumptions cannot be made about epidemiological data. Nonetheless, the effects of different cardinality limits on computational speed are explored.

The number of potential parentsets (λ) is given by 2^{μ}, where μ is the number of eligible parents. However, when a cardinality limit (c) is imposed, this reduces to:

$$\lambda = \sum_{i=0}^{c} \binom{\mu}{i} \tag{4.8}$$

Figure 4.10 shows the effect on λ of different cardinality limits for $\mu = 15$.

The influence of imposing cardinality on the dataset in the previous section (3,281 obs, 15 vars) is investigated. The cardinality limit was varied between 2 and 8, the time taken to complete 10,000 iterations (as above) is recorded in figure 4.11. The y-axis is scaled to the time taken to complete the procedures with a cardinality limit of 2 (117.8s). No limit was placed on the number of eligible parents.

The restraining of cardinality has a significant effect on the run time of the process. Again, there is a trade off between approximating the space and computational speed. In order to choose an appropriate limit, it would be necessary to perform several MCMC runs and examine the effects of imposing a limit on the results. For the analyses in this thesis, a node cardinality limit of 5 was imposed. This dramatically reduces the number of parentsets that need to be evaluated. Throughout the analyses, careful monitoring of the maximum cardinality of visited DAGs was maintained. Over all sampled network topologies, no node possessed more than 3 parents. This clearly suggests that networks containing nodes with high cardinality are low scoring, and can be reasonably ignored, making a cardinality limit

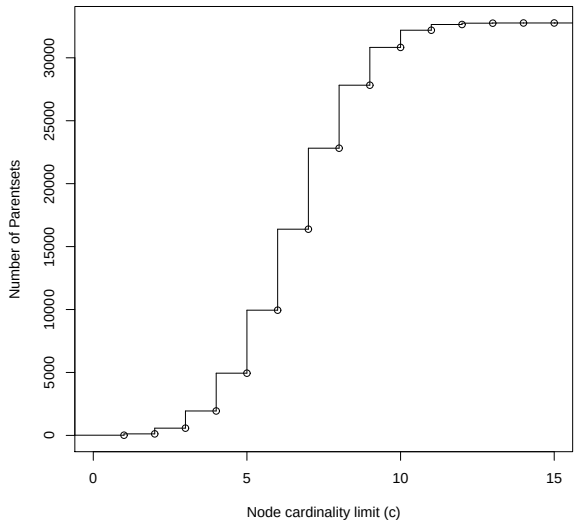


Figure 4.10: Influence of limiting node cardinality on computation time when μ=15.

of 5 a reasonable approximation. Reduced candidate sets were not imposed.

4.7 Simulated Annealing Optimisation

In some instances it is useful to identify the optimal topology with respect to the evidence criterion, possibly to ensure that out MCMC sampler is successfully traversing the best regions of topology space. Simulated Annealing (SA) is a generic optimisation technique that mimics thermal annealing by gradually reducing the probability (temperature) of a move to a lower scoring state [164]. Higher temperatures allow much more ready traversal of probability valleys. The move library described above 4.4.1 is used to move between topologies. Topologies are evaluated using the criterion described in section 3.3.1. At each step, a new DAG $\tilde{H}$ is proposed from the current DAG H. $\tilde{H}$ is accepted if υ drawn from the log uniform distribution satisfies:

$$\upsilon T < \ln \Pr(\tilde{H}|D) - \ln \Pr(H|D). \tag{4.9}$$

The value of T (temperature) is positive and gradually lowered (cooled) to 0, so the probability of accepting a lower probability DAG tends to 0. If $\tilde{H}$ has a higher probability than H then the the right hand term of eq. 4.9 will be positive, and

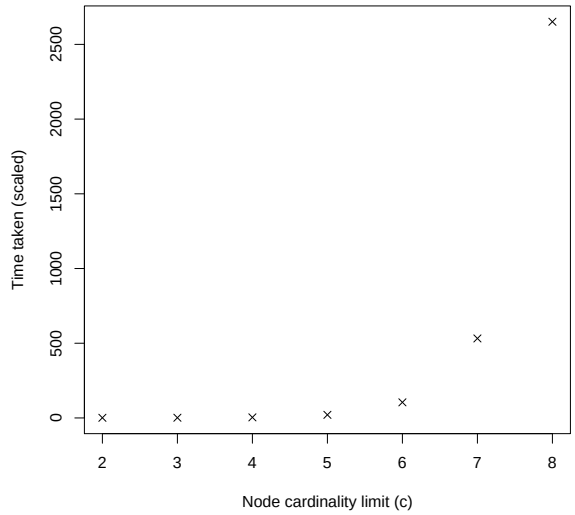


Figure 4.11: Influence of node cardinality limits on computation time

$\tilde{H}$ will always be accepted regardless of the value of T. The temperature schedule is defined by the number of temperature steps (N), the starting (T_0) and final temperatures T_N, and ω, a decay parameter. The temperature at T_n is given by:

$$T_n = T_0 - \left(\left(\frac{n}{N-1}\right)^{\omega}.(T_0 - T_N)\right). \tag{4.10}$$

The graph in 4.12 shows the temperature decay for two values of α. Unlike the Metropolis Hastings algorithm there is no need to calculate the ratio of the proposal and reversal probabilities, resulting in faster computation. The cardinality of parentsets is variable, limited to 1 more than the highest cardinality of the parentsets of all nodes, or 5 (whichever is the greater). This serves to reduce the number of parentsets that need to be evaluated by the REV move, without imposing any restrictions on the size of the search space. Without this restriction, at high temperatures the algorithm explores highly connected networks even if they score poorly. As discussed in section 4.6.4 this is very computationally expensive. The implementation of the methodology detailed in Chapter 3 is not straightforward. This chapter has discussed the difficulties associated with obtaining adequate mixing

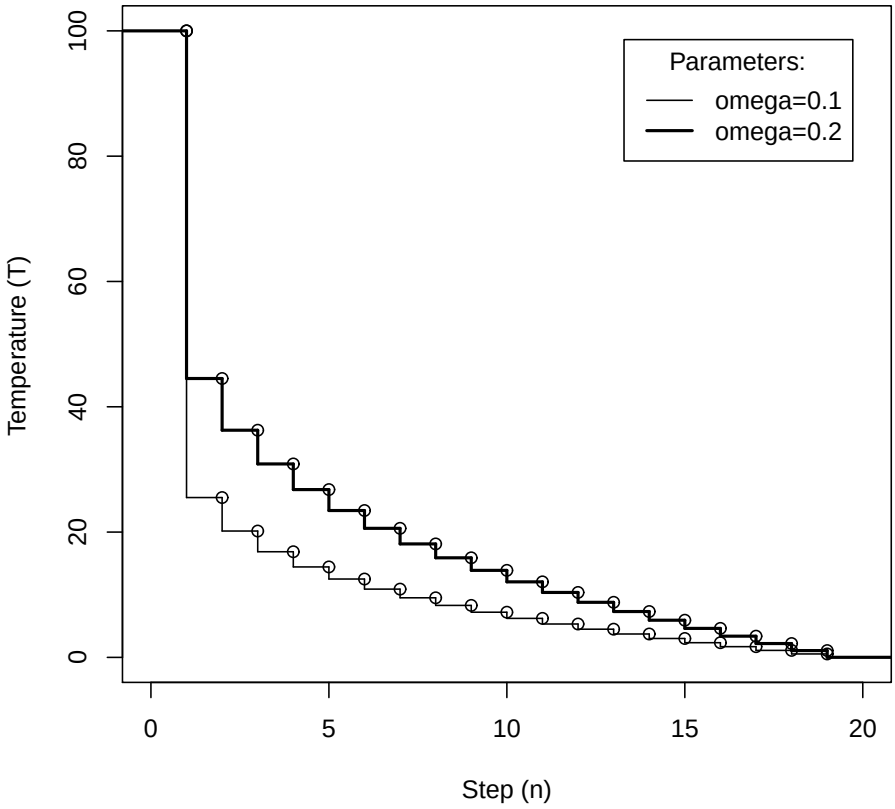


Figure 4.12: Temperature decay under different parameters.

over network structures, and tractability issues. Following this detailed grounding in the method, the thesis now continues to the three results chapters.

Chapter 5

Use of Bayesian Network Structure to Identify Factors Influencing Health Behaviour

5.1 Overview

Obesity is one of the most pressing public health concerns of the modern age. Although often oversimplified as an affliction of the 'over-malnourished' urban poor, the relationship between obesity and socio-demographic factors is complex and barely defined. In this chapter I use Bayesian networks to model obesity related factors in the 2003 and 2006 Health Surveys for England (HSE). A Metropolis Hastings algorithm is used to traverse the space of Bayesian network structures, which are scored using a criterion based on the marginal likelihood of observed data. The prevalence of particular structural features within resulting sample of network topologies is compared between male and female data. Factors influencing recreational physical activity in males and females appear to differ significantly, most notably age and education level. Relationships between social class and fruit and vegetable intake, and dietary behaviours with ethnicity and age were also observed.

5.2 Background

Obesity is a complex social phenomenon; numerous studies have reported associations between obesity-related behaviours and socio-demographic factors such as wealth, ethnicity, material deprivation and educational attainment. As discussed in the introduction to this thesis (Chapter 1) correlation between socio-demographic variables is a problem in the analysis of epidemiological datasets. Bayesian networks are an excellent tool for modelling complex intercorrelated data, and allow the investigation of all dependency relations present, rather than a single outcome variable.

In this chapter Bayesian networks are used to model interdependencies between variables in Health Surveys for England (HSE) data. The datasets contain a set of socio-demographic variables and several indicators of energy intake and energy expenditure. Interdependencies present are compared between males and females, using data from the HSE in 2003 and 2006. These surveys have been carried out annually since 1991, with the intention of providing information on the health of the nation. The 2003 and 2006 surveys focussed on cardiovascular disease (CVD) and collected detailed data on physical activity levels and diet, and consequently are of particular relevance in the context of obesity (see section 2.1 for details). Variables that describe weight status, although available, were not in-

cluded; weight status was felt to be a consequence of sustained energy imbalance as represented by the energy intake and expenditure variables present in the model. Due to the temporal nature of weight gain, adding a weight status node would require longitudinal data, as many individuals will exhibit a current energy balance not in equilibrium with their weight. The intake/expenditure variables are intended as targets for intervention, rather than predictors of BMI values.

Conditional dependencies present between variables of a dataset can be modelled by the topology of a Bayesian Network. It is possible to score the likelihood of the observed data given an specified network topology, independently of parameters (discussed in section 3.3.1). Conditional dependencies are encoded by edges (also known as arcs) between nodes. Topological information may be formalised as *structural features*; a structural feature may be the presence of an arc between two nodes, or the presence of a node within the Markov blanket of another [124]. The posterior probability of a structural feature may be evaluated following integration over the space of all possible topologies G. However, this integration is analytically intractable and exhaustive evaluation of all possible topologies is not practically possible except for very small networks. One approach is to use the highest scoring network as an approximation to G, however this is likely to exclude a significant proportion of the total probability integral. A preferable approach is to use an approximate technique that efficiently samples the space. A Metropolis-Hastings sampler is used to generate a sample $\hat{G}$ representative of the posterior distribution of network topologies, from which we can estimate the posterior probability of various structural features.

In this study we are interested in the set of posterior probabilities that each possible edge is present; *edge relation features* (ERF). The posterior probability of an edge being present between two nodes is simply estimated by counting the proportion of networks in $\hat{G}$ in which it is present. The central aim of this chapter is to use the set of edge relation features to explore, compare and contrast the relationships between obesity related variables in different populations. Differences between network topologies imply differences in the associations present in these real epidemiological systems. The edge relation features associated with a dataset can be drawn, weighting edges according to posterior probability, enabling intuitive visualisation of how the variables interact. Such observations may help to inform deeper examination of current public health initiatives as complex interventions. Further, the sampled distribution of network topologies may help shape hypotheses for obesity modelling.

This chapter proceeds as follows: Section 5.3 describes the data and the implementation. Section 5.4 provides the results of the Metropolis Hastings sampler. The final section provides a discussion of the findings.

5.3 Approach and Methods

5.3.1 Data

Two panels of data were considered by sex; males and females, from 2003 and 2006 HSE data. Individuals under 16 years of age were excluded from the study; children's eating and exercise habits are strongly influenced by parental behaviour [165, 166], questionnaire contents were different, and it was felt that they were not suitable for inclusion in the model. Individuals over 74 were excluded as studies have shown that overweight and obesity is less of a risk factor for morbidity and mortality in the elderly [167], thus they are not the main targets for intervention. Regrettably, due to differences in questionnaire structure between CORE 1 and CORE 2 in 2006, it was also necessary to exclude those over 65 in CORE 1 from the 2006 data (see 2.1). The 2006 HSE surveyed 21,399 individuals. The following individuals were excluded; those aged under 15 or over 74 years (8,759), CORE 1 individuals aged over 65 (907), individuals failing to fill in the self completion (SC) booklet (3,075) and those missing other variables present in the model (139). The final dataset consisted of 8,159 individuals; 3,806 males and 4,713 females. The 2003 survey contained 18,533 individuals. Following exclusion of individuals outside the age range (5,177), those failing to fill in the SC booklet (3,088) and those with other missing values (59), the final dataset consisted of 4,572 males and 5606 females (10,178). To enable comparability between years, variables are kept consistent between annual health surveys where possible. Consequently, equivalent variables were available between all datasets, with the exception of Incidental Physical Activity (IPA) which due to different questionnaire structure is slightly different between the two datasets (2.1.3). The variable set using the definitions provided in section 2.1.3 and 2.2.3 is listed in table 5.1.

5.3.2 Metropolis Hastings Sampling

Metropolis-Hastings (MH) sampling is an MCMC technique used to generate a sample representative of the posterior distribution of topologies of Bayesian net-

Socio-Demographic Variables	
Variable	**Abbreviation**
Sex	*Sex*
Age (groups)	*Age*
Dependent Children	*Dep.Cld*
Marital Status	*Marital S*
Health Status	*Health*
National Statistics Socio-Economic Classification	*NS-SEC*
Economic Status	*Econ.S*
Ethnicity	*Ethnic.*
Education Level	*Educ.L*
Transport Access	*Trans.A*
Leisure Access	*Leis.A*

Energy Expenditure Variables	
Variable	**Abbreviation**
Recreational physical activity level	*Rec.PA*
Incidental physical activity level	*Inc.PA*
Occupational physical activity level	*Occ.PA*

Energy Intake Variables	
Variable	**Abbreviation**
Fried food intake level	*FriedFd*
Cake/sweets intake level	*Cakes*
Snack/crisps etc. intake level	*Snacks*
Fruit and vegetable intake level	*Frt Veg.*

Table 5.1: List of Health Surveys for England variables used in this analysis; identifying factors associated with health behaviour

work models of the data. In the Markov Chain process, the probability of acceptance of a new topology H^{t+1} is conditional on the ratio of the evidence for the topologies of H^{t+1} and H^t, subject to the requirements of detailed balance. Consequently topologies with the highest score are preferred and best represented within the sample. The sample $\hat{G}$ allows us to the estimate the posterior probability of various structural features using Bayesian Model Averaging (BMA). Metropolis-Hastings sampling, BMA, and topology scoring are all discussed in detail in chapter 3. For each dataset, four independent runs of the sampler were performed. Two runs were initialised from an empty network, and two from the optimal network

as discovered using a simulated annealing optimisation algorithm. These runs are referred to as *empty* and *informed* respectively. As outlined in section 4.3, multiple initialisations were carried out in order that the mixing and convergence of the chains can be monitored.

In each run, a burn-in period of 5×10^5 iterations was performed, before a sampling period of a further 5×10^5 iterations. Samples were taken every $1,000$ iterations to provide a final sample $\hat{G}$ of 500 topologies. The value of the pseudo-counts or hyperparameters (α_{ijk}) was set at 1.0.

The moves used to propose network topologies are described in section 4.4.1, and were used in the following ratio, determined following analysis of mixing properties:

- Add Arc: 0.4.
- Remove Arc: 0.4.
- Grzegorczyk-Husmeier REV move: 0.1.
- Switch Arc: 0.05.
- Multiple Reversal move: 0.05.

The edge relation features are estimated by counting the proportion of network topologies in which the edge is present, the orientation is ignored. The set of edge relation features is generated by determining the posterior probability of edges for each pairwise combination of nodes.

5.4 Experimental Results

Results are shown for each of the four datasets. Edge relation features (ERFs) are displayed in graphical form, arcs are undirected as arcs in both directions are included in my definition of a feature. Arcs are colour coded to minimise visual clutter: black indicates the arc was present in every observed topology, red if present in 0.1 or greater of topologies and orange if present in less than 0.1. More sophisticated colour coding systems were considered, but felt to add little value following inclusion of the ERF values as labels.

In addition to edge relation features, panels comparing the results of the different initialisations are displayed. These scatter diagrams plot the edge relation features of a) the two empty initialisations (empty-empty), b) the two informed

initialisations (informed-informed) and c) the empty and informed initialisations (empty-informed). These serve to check the successful convergence of the MCMC runs, where mixing and convergence is high, agreement will be high amongst all initialisations, methods of evaluating mixing are discussed in section 4.3. Where empty and informed initialisations are plotted against each other runs that failed to converge are excluded, this is denoted by an asterisk (*). Edge prevalences are included as labels on arcs. The optimal topology for each dataset is included for completeness (figures 5.5, 5.6, 5.7 and 5.8). Where large volumes of data generate a highly peaked topology space G, the optimal topology will tend to be similar to the resulting Metropolis-Hastings sample [168].

5.4.1 Males: 2006 data

Results from the Metropolis Hastings sampling over the 2006 male data are displayed in this section. One of the empty chains did not converge successfully, as shown by the evidence traces in figure 5.9. The other empty chain successfully converged with the two informed initialisations. Interestingly, the scatter plots of empty vs. empty edge relation features in figure 5.10 show almost perfect alignment with minimal indication of poor convergence. This suggests that the space explored by the non converging chain, although distinct and less high scoring than the space of the other chains, appears equivalent in terms of the edge relation features. This implies the DAGs explored by the non converging chain had similar structure with a few arc reversals. However the chain was unable to reach the highest scoring region over the sampling period. Despite its equivalence, this run is excluded when calculating the edge relation features.

Figure 5.1 is a graphical representation of the edge relation features observed. We can begin to make some interesting observations about the underlying structure of the data.

Age is highly connected, displaying conditional dependencies with numerous other variables, unsurprisingly with socio-demographic variables, but also behavioural variables such as recreational physical activity levels (RPA) and snack consumption. RPA also exhibits a strong conditional relationship with health status.

Occupational physical activity (OPA) shows associations with Economic status, Social class, Education Level and Health. In all sampled networks arcs were present between these four variables and OPA. Incidental physical activity (IPA), *i.e.* walking behaviour, is influenced by Health and Education level. The ERF

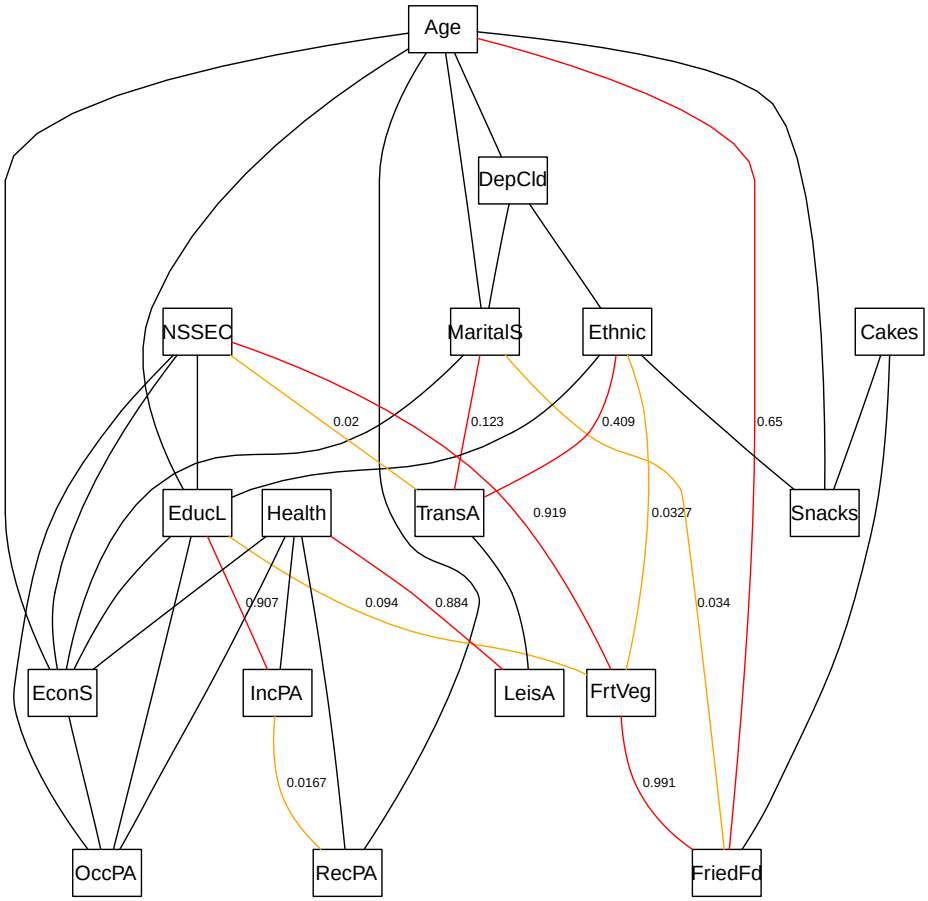


Figure 5.1: Relationships between eating, physical activity and socio-demographic factors in males presented as the average Bayesian network topology, derived from Metropolis Hastings sampling (2006 data). Arcs colour coded: black, arc present in all observed topologies; red, ≥0.1; orange, ≤0.1

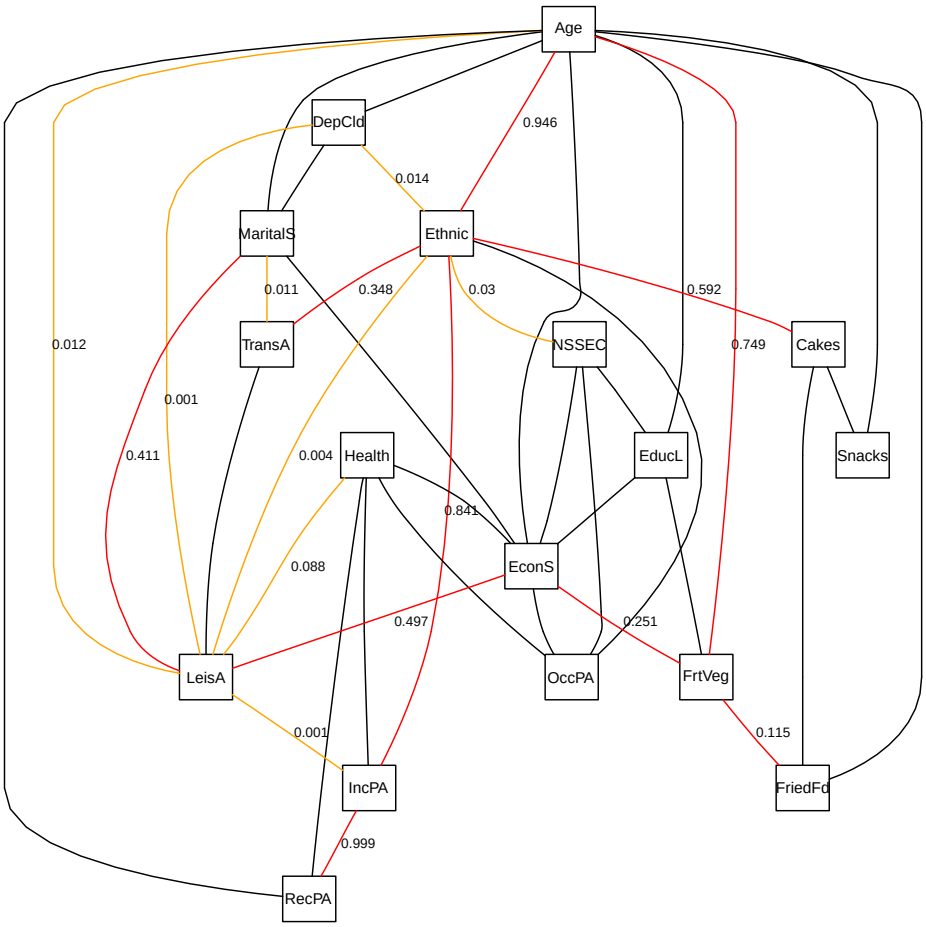


Figure 5.2: Relationships between eating, physical activity and socio-demographic factors in males presented as the average Bayesian network topology, derived from Metropolis Hastings sampling (2003 data). Arcs colour coded: black, arc present in all observed topologies; red, ≥0.1; orange, ≤0.1

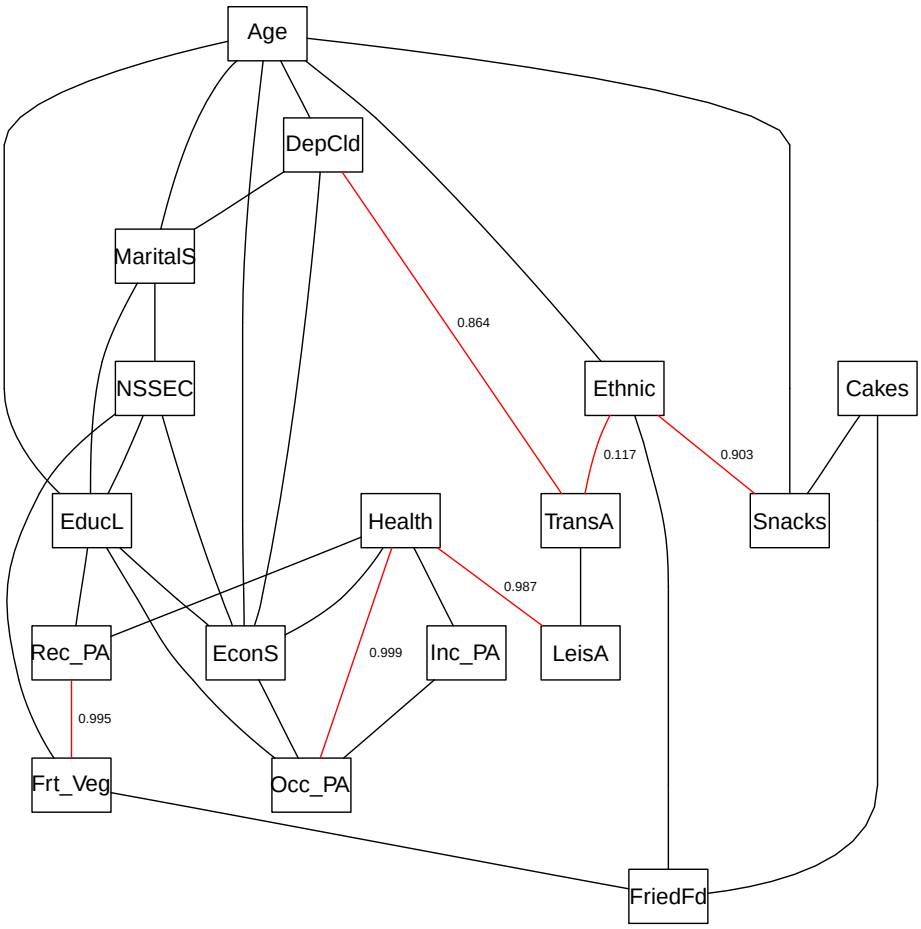


Figure 5.3: Relationships between eating, physical activity and socio-demographic factors in females presented as the average Bayesian network topology, derived from Metropolis Hastings sampling (2006 data). Arcs colour coded: black, arc present in all observed topologies; red, ≥0.1; orange, ≤0.1

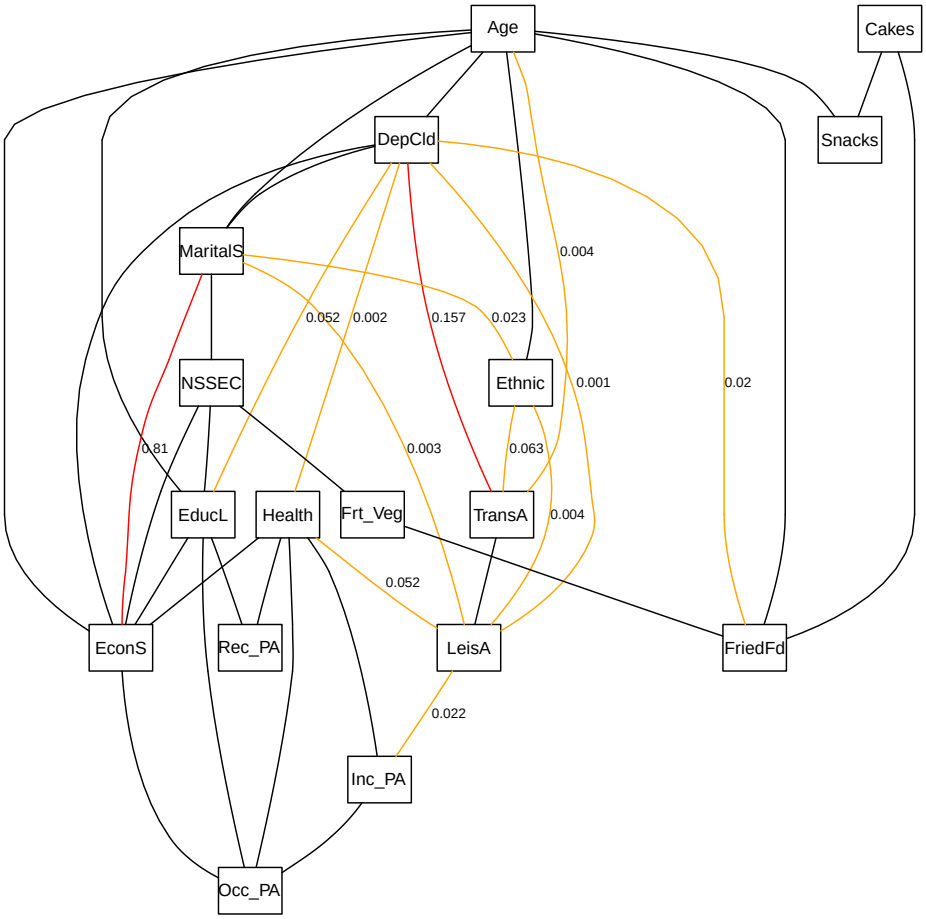


Figure 5.4: Relationships between eating, physical activity and socio-demographic factors in females presented as the average Bayesian network topology, derived from Metropolis Hastings sampling (2003 data). Arcs colour coded: black, arc present in all observed topologies; red, ≥0.1; orange, ≤0.1

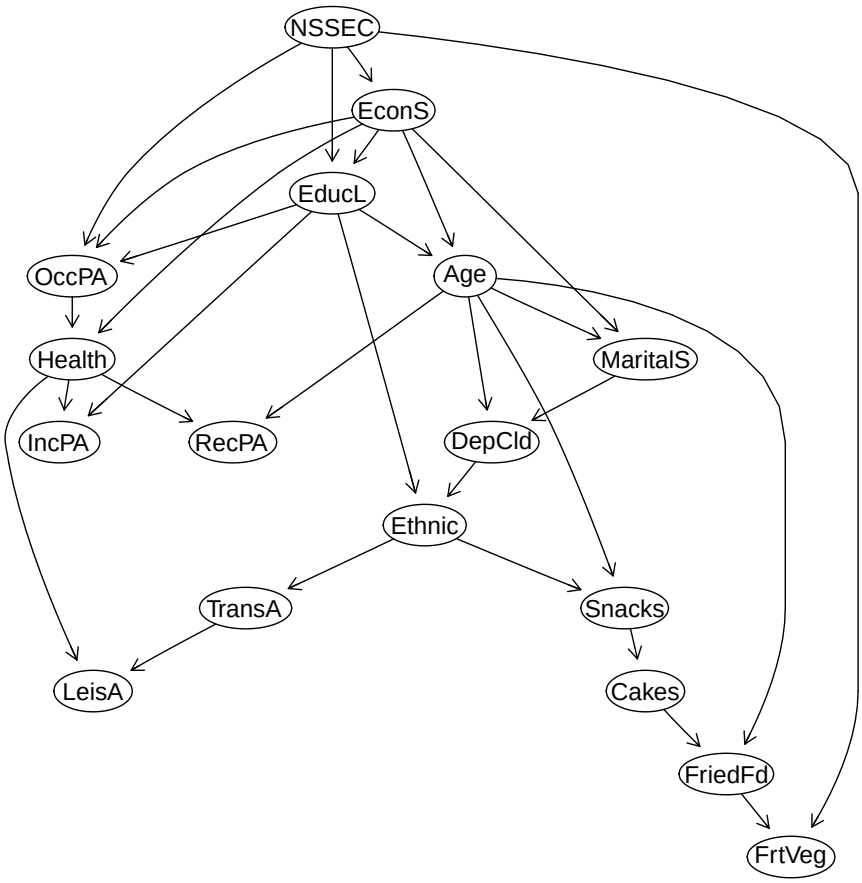


Figure 5.5: Optimal Bayesian network topology of obesity related factors from Health Survey for England data discovered using simulated annealing (Males 2006)

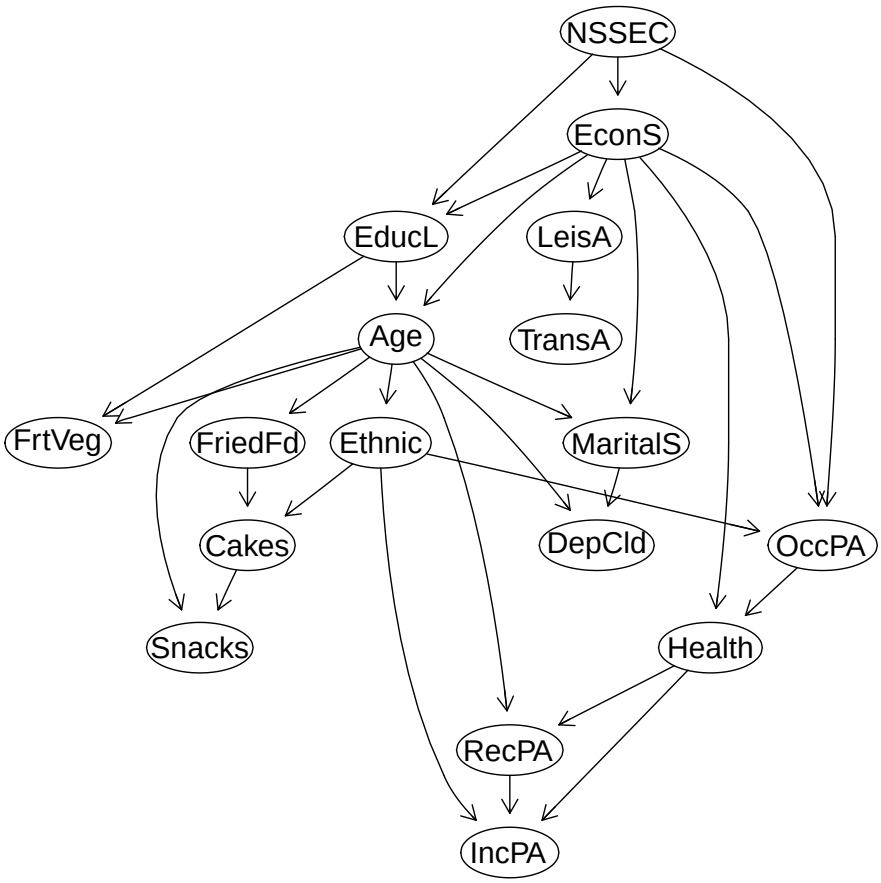


Figure 5.6: Optimal Bayesian network topology of obesity related factors from Health Survey for England data discovered using simulated annealing (Males 2003)

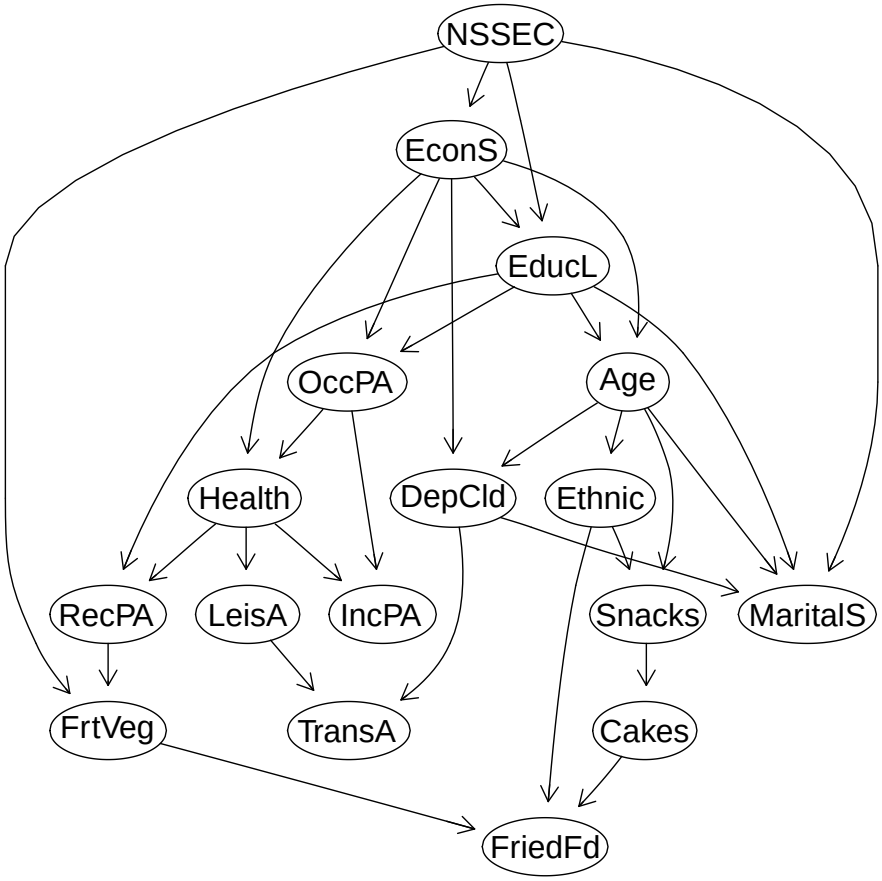


Figure 5.7: Optimal Bayesian network topology of obesity related factors from Health Survey for England data discovered using simulated annealing (Females 2006)

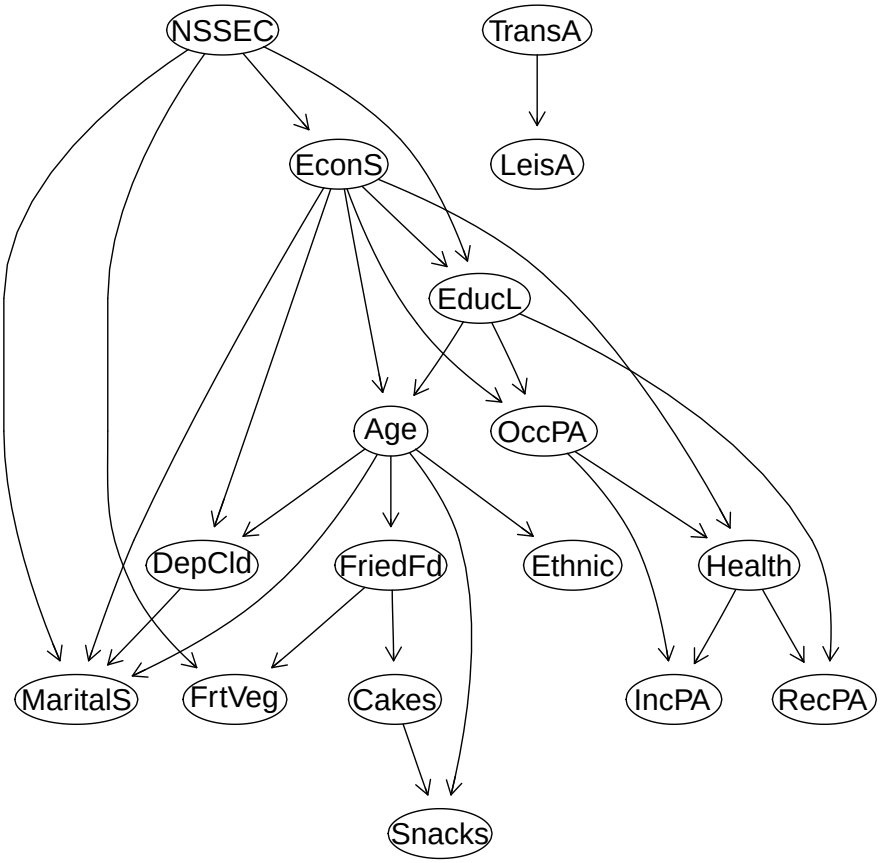
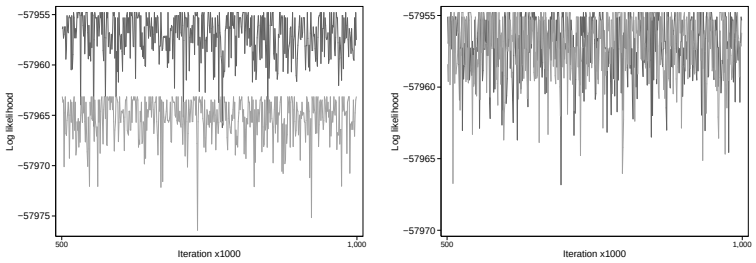
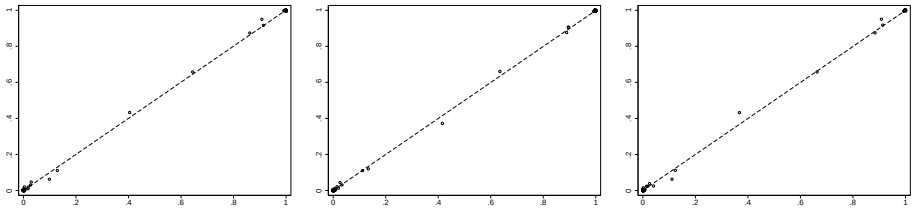


Figure 5.8: Optimal Bayesian network topology of obesity related factors from Health Survey for England data discovered using simulated annealing (Females 2003)



(a) empty/empty (b) informed/informed

Figure 5.9: Evidence traces of Metropolis Hastings sampling process (male 2006 data)



(a) empty-empty (b) informed-informed (c) empty*-informed

Figure 5.10: Scatter plots of edge relation features obtained following Metropolis Hastings sampling (male 2006 data)

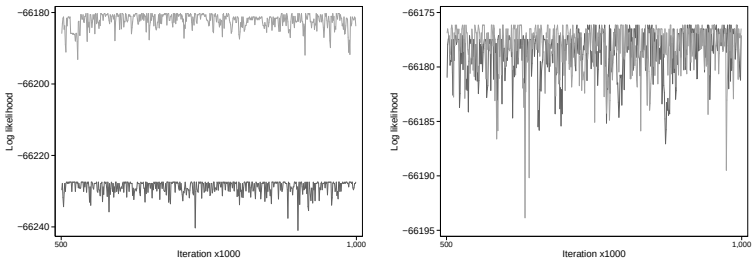
between Education and IPA is 0.91 in contrast to the 1.0 of edge relation features previously described. This indicates that slightly lower confident in the presence of an arc here, though it still suggests an important correlation.

The energy expenditure variables exhibit a high degree of intercorrelation. Ethnicity exhibits conditional dependence with snack food intake, and (weakly) with Fruit and Vegetable consumption. Age also appears to be an important explanatory factor behind dietary intake. Social class and Education display correlation with Fruit and Vegetable intake.

5.4.2 Males: 2003 data

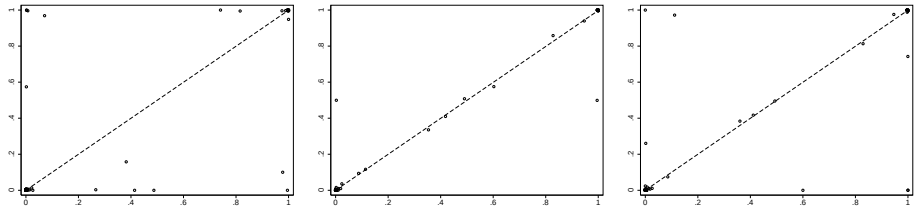
Both empty initialisations failed to converge to the most probabilistically dense region of topology space (figure 5.11). The informed initialisations also experienced some problems, one of the initialisations spending some time exploring a distinct, but still high scoring region. There appear to be two regions of similar probability,

and transition between the two is difficult. The two regions appear to differ by a single arc, with the highest scoring region having an edge between Economic status and leisure access, compared to between marital status and leisure. The peak of the former region is approximately 2.7 times more likely than that of the other region. As transition is so rare, a very large number of iterations would be required for the sample to accurately represent the relative probability of the two regions. To incorporate this uncertainty, both runs are included when calculating the ERF set.



(a) empty/empty (b) informed/informed

Figure 5.11: Evidence traces of Metropolis Hastings sampling process (male 2003 data)



(a) empty-empty (b) informed-informed (c) empty*-informed

Figure 5.12: Scatter plots of edge relation features obtained following Metropolis Hastings sampling (male 2003 data)

The topology distributions for 2003 males show similar results to the 2006 data. RPA is consistently linked with age and health, however there is also a consistent edge with incidental physical activity (IPA). IPA also appears to have some dependency with ethnicity. Occupational physical activity has similar interdependencies, although ethnicity is again included. Age is again associated with dietary intake variables, which are closely correlated. Fruit and vegetable intake is linked with

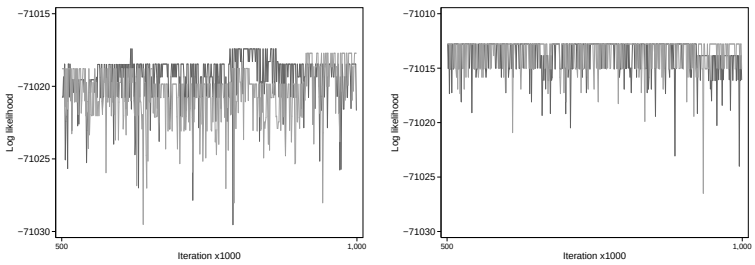
education level rather than social class. Ethnicity appears to have some correlation with dietary intake variables as observed in the 2006 data.

Despite some differences, the overall interdependencies present in the data remain fairly similar between the 2003 and 2006 HSEs. Due to sampling variation, a higher proportion of individuals of a certain age, social group or ethnicity may result in changes to the overall topology. Expectation of a greater level of agreement is unreasonable.

5.4.3 Females: 2006 data

The Metropolis Hastings algorithm encountered more severe mixing issues with the female 2006 data than observed previously; presumably due to the larger dataset resulting in a more jagged probability distribution [168] and thus having a greater tendency to become stuck in local maxima [124]. Neither of the empty initialisations converged on the space explored by the informed initialisations (figure 5.13), agreement of edge relation features between these runs was poor (fig. 5.14).

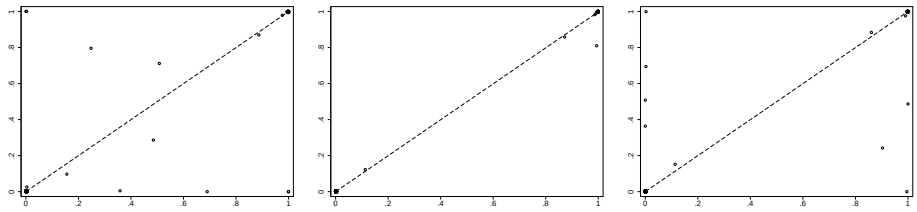
In the informed initialisations, the evidence traces in figure 5.13(b) show that the chains explored similar space until one chain moved to a less high scoring region at approximately 9.2×10^5 iterations. This is the cause of the slightly anomalous edge relation feature in figure 5.14(b) (data not shown), this is an unavoidable result of the large dataset pending further heuristic improvements. To estimate the edge relation figures empty initialisations are excluded.



(a) empty/empty (b) informed/informed

Figure 5.13: Evidence traces of Metropolis Hastings sampling process (female 2006 data)

As the probability distribution is so peaked, the distribution of edge relation features becomes highly bimodal, the probability of each feature approaches 0 or 1. This reflects our increasing confidence in the topology with increasing data.



(a) empty-empty (b) informed-informed (c) empty*-informed

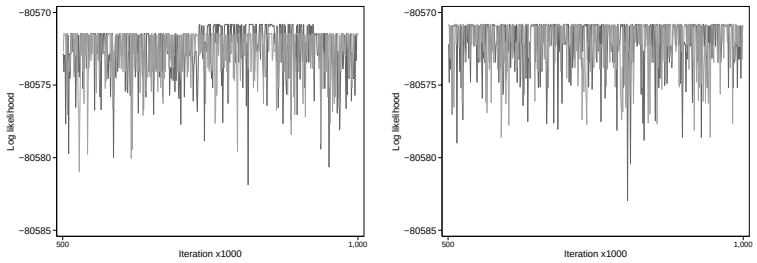
Figure 5.14: Scatter plots of edge relation features obtained following Metropolis Hastings sampling (female 2006 data)

In females similar high levels of correlation within socio-demographic variables and dietary intake variables are observed. RPA is associated with education level rather than age, interestingly RPA also exhibits strong conditional dependency with fruit and vegetable intake, this may reflect a social pattern, or a latent tendency for health conscious women to eat more fruit and vegetables and take regular exercise. Incidental physical activity (IPA) is consistently associated with health status, as in males. There is also a suggestion of association between IPA and other occupational activity, though this may be a result of retired women walking more. Specific interdependencies are examined in more detail later. Relationships between social class and fruit and vegetable intake, and ethnicity and dietary indicators are again observed.

5.4.4 Females: 2003 data

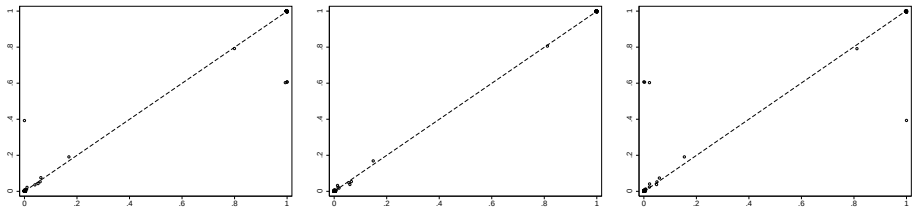
Although the female 2003 dataset is the largest of the datasets examined, reasonable mixing was observed. One of the runs from an empty initialisation failed to converge on the optimal topology completely, while the other did eventually visit after 7.0×10^5 iterations, leaving the high scoring region soon afterwards (fig. 5.15). The informed topologies explored the space surrounding the peak topology and generated equivalent results (fig. 5.16).

Figure 5.4 displays the edge relation features derived. The estimates of edge posterior probabilities shows a great deal of similarity with the female 2006 data. The majority of the few edges that differ between the two datasets involve ethnicity. Ethnicity is much less connected in the 2003 data, possibly due to the smaller proportion of non-white individuals in this study (7.0% vs 8.8%). The RPA-Fruit and Vegetable intake arc, present in every topology in the 2006 data, was not observed



(a) empty/empty (b) informed/informed

Figure 5.15: Evidence traces of Metropolis Hastings sampling process (female 2003 data)



(a) empty-empty (b) informed-informed (c) empty*-informed

Figure 5.16: Scatter plots of edge relation features obtained following Metropolis Hastings sampling (female 2003 data)

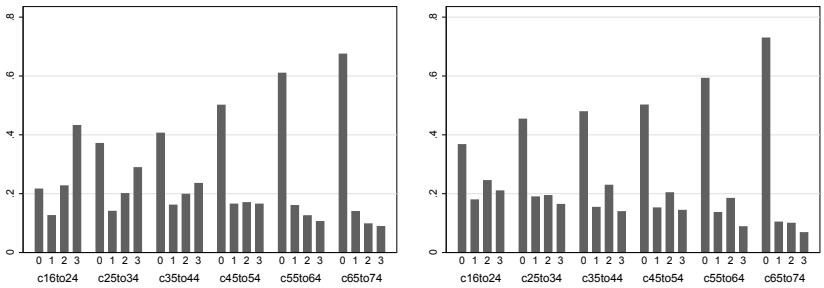
at all in the 2003 data. This may be due to a closer relation between NSSEC and Fruit and Vegetable intake resulting in less need for further explanatory variables.

5.5 Further Analysis and Interpretation of Results

The aim of this study is to identify factors that influence health behaviour, consequently I focus on potential determinants of energy intake and expenditure variables. The sampled topologies derived from Metropolis Hastings sampling provide information regarding the conditional dependencies present in the data. In this section the findings are explored in more depth, focusing on the more recent HSE 2006 data.

Gender Differences in Determinants of Recreational Physical Activity

An interesting observation from the discovered network topologies is the differs conditional dependencies that recreational physical activity (RPA) exhibits between males and females. According to the sampled topologies, in males RPA has a dependency with age and health. In females we observe a relationship with education level, health and fruit and vegetable intake. Here the overall topology becomes important; females education level is a parent of RPA, and has age as a parent, implying conditional independence of age and RPA (see section 3.1.2). This structure is observed in all sampled topologies. Given this topology we may expect female RPA levels to display correlation with age, as shown in figure 5.17, but to be independent of age given education. The bar graphs in this section display the maximum likelihood estimate of the probability of falling into each of the 4 RPA categories, given parents. Health is excluded by only including those in good health in these graphs.

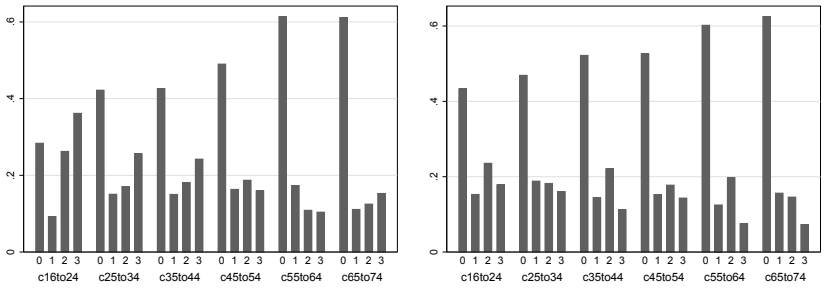


(a) Males (b) Females

Figure 5.17: Probability estimates of recreational physical activity behaviour categories by age group

The structure of the discovered topologies implies that female RPA is independent of age given education. To explore this I fix on education, including only those in the 'Below Higher Education' category (as it is the most common). Given the topologies observed, we expect males to show substantial variation between age categories and females little. Figure 5.18 shows the relevant maximum likelihood probabilities.

The age shift of RPA appears significantly less pronounced in females than males, the conditioning over education level has had a significant reduction on the influence of age. However, age does appear to have some residual influence. This

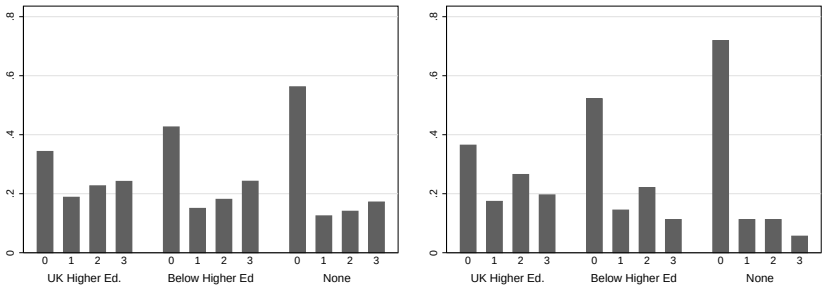


(a) Males (b) Females

Figure 5.18: Probability estimates of recreational physical activity behaviour categories by age group of individuals in education level category 'below higher'

highlights a weakness of the knowledge discovery technique; a lack of sensitivity is discussed later. It is worth noting that the graph does not display the relative numbers of individuals in each group or the associated uncertainty.

Given the observed topologies we expect males to be independent of education given age. Here I fix at age '35-44' (again the most numerous category) and plot results for education level as a parent (figure 5.19). The 'current student' group is excluded as it is rare in this age category.



(a) Males (b) Females

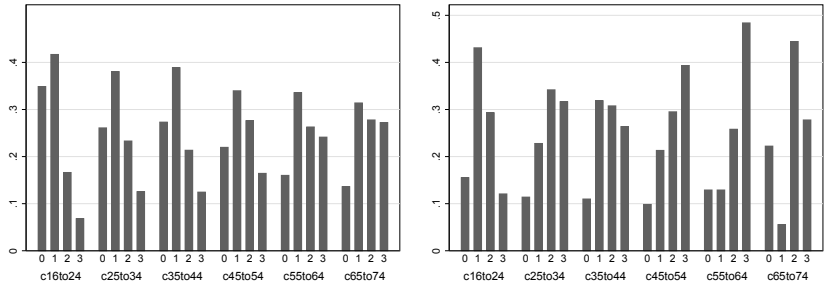
Figure 5.19: Probability estimates of recreational physical activity behaviour categories of individuals in age group '35-44'

Smaller differences between educational categories in male 35-44 year olds are observed than those in the female data. The method implemented here may have identified different dynamics of recreational physical activity in males and females which may be of interest to policymakers. It should be stressed that these bar

graphs only represent a small proportion of the observed data.

Influence of Ethnicity and Age on Snacking Behaviour

In both the male and female topology sample we observe consistent edges between ethnicity and age with snacking behaviour. Dietary variation has previously been noted between different ethnic groups, but has been difficult to separate from other social indicators. In males (figure 5.20) we observe radically different patterns of reported snack food consumption by ethnicity. The non-white group exhibits a far lower proportion of individuals in the highest consumption group, and is broadly consistent between age categories, although there may be a pattern of increasing consumption with age. In contrast the white group shows a very high level of reported snack food intake in younger individuals and a dramatic decline with increasing age. Ethnic differences are less pronounced in females (figure 5.21), with the non white group reporting slightly higher snack intake in younger individuals. This may reflect a real behavioural difference, or ethnic variation in reporting bias. Comparison between ethic groups is limited by the small numbers of non white individuals available, particularly in older age groups (table 5.2).

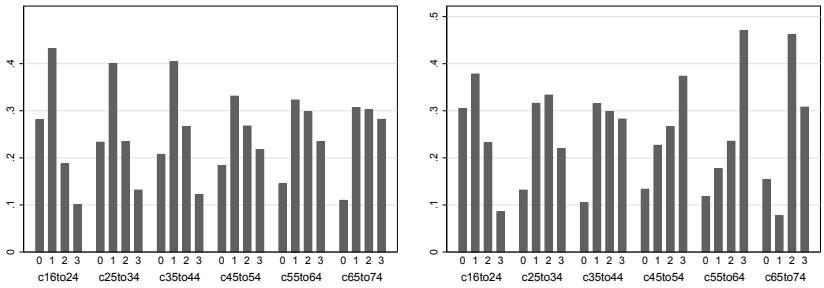


(a) White (b) Non-White

Figure 5.20: Probability estimates of snack consumption category by age group, males.

Fruit and Vegetable Intake by Social Class

Social class has frequently been reported as a determinant of fruit and vegetable intake, the network discovery technique implemented here also identified this association [57, 61, 169]. Figure 5.22 displays the maximum likelihood probability estimates of fruit and vegetable consumption categories by social group. The data



(a) White (b) Non-White

Figure 5.21: Probability estimates of snack consumption category by age group, females.

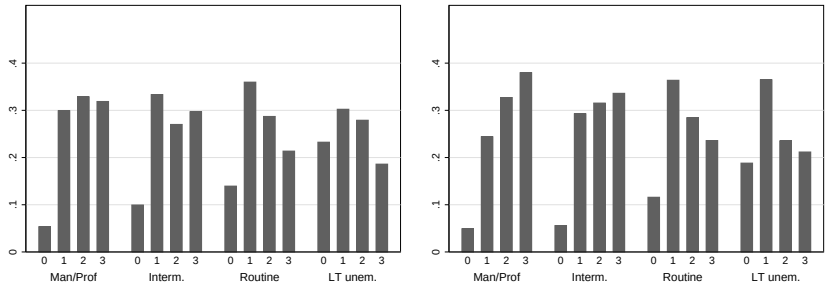
Age	Sex	White	(%)	Non-White	(%)	Total
16-24	Male	363	(87.1)	54	(12.9)	417
	Female	445	(85.1)	78	(14.9)	523
25-34	Male	498	(86.9)	75	(13.1)	573
	Female	665	(85.8)	110	(14.2)	775
35-44	Male	783	(90.0)	87	(10.0)	870
	Female	1023	(89.5)	120	(10.5)	1143
45-54	Male	720	(92.7)	57	(7.3)	777
	Female	878	(92.5)	71	(7.5)	949
55-64	Male	833	(96.9)	27	(3.1)	860
	Female	950	(96.9)	30	(3.1)	980
60-74	Male	356	(96.2)	14	(3.8)	370
	Female	397	(97.8)	9	(2.2)	406
Total	Male	3553	(91.9)	314	(8.1)	370
	Female	4358	(91.2)	418	(8.8)	406

Table 5.2: Counts of individuals in 2006 HSE data by sex and ethnicity

clearly suggests males and females in lower social classes consume fewer fruit and vegetables. The conditional dependencies identified by the technique described do not imply causality, essentially the the interactions seek to build the network topology that best explains the observed correlations. The main aim of the method is to generate hypotheses by identifying conditional dependencies in complex datasets.

Some interesting observations can be made from the presence of network structural features:

- Age is strongly associated with recreational physical activity levels in males,



(a) Males (b) Females

Figure 5.22: Probability estimates of fruit and vegetable intake by social class

in females educational attainment may be a better predictor.

- Ethnicity is consistently related to dietary factors in males and females.
- Age is associated with snacking behaviour and other dietary intake variables in both males and females.
- Social class is correlated with fruit and vegetable intake in both males and females.

5.6 Discussion

This work represents a novel application of an MCMC Bayesian network sampling algorithm to a real epidemiological dataset. This approach has the advantage of coping well with highly correlated covariates which are an inherent feature of socio-demographic data. States of numerous nodes can be modelled simultaneously as a set of complex relationships rather than just the outcome of a specified dependent variable. Identification of common structural features provides insights into the interdependencies present within the data. Results are easily transformable into a graphical output which enable a user to see the complex relationships present. This approach has the ability to highlight probabilistic relationships between variables without concerns of investigator bias, and represents a useful tool for hypothesis generation.

The most striking finding of this analysis is the markedly different dynamics of RPA in men and women. A significant component of the influence of age on recreational physical activity in women is explained by the association between

age and education level. This implies that education level, not age is the primary determinant of recreational physical activity in women. In males, although age and education level are correlated, age is the best predictor of recreational physical activity habits. Traditional methods examining similar data [170] failed to uncover such conditional independence relationships, highlighting the advantages of this approach. An association between social class and fruit and vegetable intake was also highlighted, as was variation in snack food intake by ethnicity and sex. Crucially, these observations were not based on *a priori* hypotheses, but generated by a **data-driven** approach. Machine learning is seldom applied in epidemiology at present, but it represents a powerful tool for identifying dependence relationships.

It is important to bear in mind what the topology discovery method is doing, which is explaining correlations in terms of network structure. The method attempts to approximate the posterior distribution of network topologies G given a scoring criterion, networks are evaluated given data as a whole. Consequently, the best explanation of a specific node is not necessarily consistent with that of a network. A strong effect in a small group or a relatively weak effect may not influence the evidence for the model sufficiently to result in an edge. An example of this was observed in the residual effect of age on female RPA levels, when education level was fixed (section 5.5). Nonetheless, the technique is well suited to highlighting dependencies in complex datasets.

The high level of agreement between the two years of the HSE is encouraging, as it suggests that the structural features observed within the data are genuine rather than artefacts of the data. Differences in network topologies between the 2003 and 2006 data may be explained by sampling variation, a population with a higher proportion of a specific group will reflect this in the high scoring network topologies. For example if social class is only important in determining the fruit and vegetable intake of the young, then an older population would be much less likely to display the relevant edge. Robustness of results is something that requires further investigation. In a peaked dataset, edge relation features are not greatly informative, as they tend to become binomially distributed around 0 or 1. Bootstrapping would indicate how sensitive our topology distributions are to sampling variation and even enable the construction of confidence intervals. However, as the main aim of this work is hypothesis generation, rather than accurate estimation of edge relation features, this was felt to be of limited value.

Identification of conditional dependencies may help to inform future interventions. In men we see a strong decline in RPA with age (fig. 5.17(a)); policy makers

may therefore wish to concentrate on initiatives to encourage men to remain active as they get older. In contrast RPA levels in women are much more dependent on social factors, best explained by education level. Interventions to increase RPA amongst women may focus on financial and physical accessibility and other social issues. Social class is a strong determinant of fruit and vegetable intake in both men and women, in agreement with various studies [57, 62], as noted previously this does not imply a causal relationship, but does identify which groups are eating less healthily. More research is needed to identify potential interventions, but should focus on those of lower socio-economic status. A relationship was observed between ethnicity and dietary factors. These effects may reflect cultural differences that may be a potential target for policymakers.

Although applying Bayesian networks in this manner is an effective means of reducing bias due to correlation between predictor variables, the data is still subject to limitations. The sample is not fully representative of the UK population due to participation bias, thus we must exercise caution when making inferences about the UK population. Exclusion of individuals due to missing data points is likely to exacerbate this bias. Other biases are also likely to be present in survey data; self reporting of food intake and physical activity levels may not be accurate, and inaccuracies may vary across population groups. Some of the variables included in the model may not accurately reflect the true status of the individual, classifications may be inaccurate or insufficient. The failure to accommodate continuous data represents a loss of information, which reduces the sensitivity of this analysis, which is discussed in more detail later (Chapter 8).

The methodology described in this paper assumes that the sample ($\hat{G}$) of the space of all possible DAG topologies (G), represents the vast majority of the total integral. If the sampling algorithm has difficulty mixing this assumption may not hold. The approach taken here of initialising runs with the optimum topology is only valid when this region contains the vast majority of the total probability integral. As the distribution of network topologies approaches a Dirac delta distribution as the number of observations tends towards infinity, this is reasonable for large datasets. The datasets used in this study probably reach the limit of the usefulness of Metropolis-Hastings sampling as a method to approximate G due to poor mixing. Beyond this, a strategy based on searching for unique high scoring topologies may be more fruitful. Alternatives are discussed in the concluding chapter.

The addition of more variables could extend the method, imposing little ex-

tra computational load. However, it would result in a more complex topology space and thus more difficult mixing. MCMC over network topologies is an active research field, and further developments may allow use of larger more complex datasets. Other possible extensions of the method include the introduction of missing data, and the possibility of latent variables. Although straightforward to include as an extra outcome within each variable, missing data was not included in this study, as edges would seek to explain correlations between social factors and nodes with a high proportion of observations missing.

The derived sample of network topologies has further utility, once parameters have been learned 3.3.2. Chapter 6 uses a complete Bayesian network model to make predictions regarding population health behaviour. The technique described here represents a method of identifying conditional dependencies in complex discrete datasets. In carrying out this work, I have developed a general toolkit that can perform MCMC sampling and optimisation of Bayesian network topologies which may be applied to other datasets.

Chapter 6

Combination of High and Low Resolution Datasets Using Bayesian Networks

6.1 Overview

In the UK, population estimates of levels of obesity related behaviour are scarce. Although the Health Surveys for England collect relevant data for a large number of individuals, only summary statistics are reported [170]. Due to participation bias, the utility of Health Surveys for England (HSE) data is limited as it is not representative of the UK population. Furthermore, local populations are demographically distinct, and are likely to exhibit different health behaviours. This analysis utilises Bayesian networks to combine a small high resolution dataset of health behaviour with a larger lower resolution dataset in order to estimate health behaviour at a population level. A Bayesian network model of energy intake and expenditure behaviours given a range of socio-demographic factors is constructed from HSE data. Data from the 2001 UK census is then applied to this model to estimate obesity related behaviours in Greater Manchester.

6.2 Background

Obesity is a major cause of chronic disease in the UK [15] which represents a significant proportion of health expenditure, both directly and indirectly. Currently, despite a need for action being identified by the Government in 2001 [15], there is a lack of evidence for the effectiveness of policy interventions. This is in contrast to direct medical interventions such as drug treatment and bariatric surgery, which have been the subject of numerous trials [87, 171, 172].

The obesity epidemic is most likely due to a shift in population behaviour over the past few decades. A significant stumbling block for policy interventions is the lack of population data on energy intake (EI) and energy expenditure (EE) behaviours [8], such data is necessary to inform policy interventions. There are inherent difficulties associated with estimating population behaviour. Direct measures of energy expenditure are available using techniques such as doubly-labelled water; however, they are expensive and impractical for large studies. Surveys are currently the only viable method of obtaining relevant data. Although survey data can be inaccurate [131, 132], it provides an effective method to collect the large amounts of data required.

Population health behaviours are frequently correlated with socio-demographic factors such as social class, education level, and wealth [57, 59, 68, 73, 173]. Geographic areas are demographically distinct even at higher levels such as wards

or metropolitan districts. Consequently, we can reasonably expect that such areas will have different population behaviours owing to their different demographic characteristics.

A number of studies have attempted to identify determinants of physical activity [174–177] (particularly of adolescents [178–181]) and diet quality [57–61]. However, there are a dearth of studies that attempt to estimate levels of energy intake or expenditure over an entire population. The UK Diet and Nutrition Survey [182] has provided some detailed information on dietary intake, and is set to report energy expenditure data from doubly-labelled water studies. However, the small sample size and high attrition rate diminish its usefulness.

The Health Surveys for England (HSE) are the primary information source for population health behaviours in the UK; Stamatakis *et al* have previously analysed physical activity data from the HSE, reporting temporal trends [183, 184] and summary statistics [170]. The data in the HSE cannot be representative of a specific population, as it is a national study with known issues of participation bias. As obesity policy is typically implemented by Local Authorities and Primary Care Trusts, it is valuable to have data reflecting the behaviour of a local population.

The current chapter describes a radically different approach to previously analysed data. A model of energy intake and expenditure indicators is constructed from HSE data; this model is applied to the socio-demographic characteristics of a real population. The resulting data represents a synthetic estimate of the obesity related behaviour of this population. This information might be useful to local policy makers.

The model implemented is a multi-output Bayes classifier learned from HSE data. An advantage of a methodology based on Bayesian networks is the ability to simultaneously model multiple outputs that are likely to be correlated due to the known effect of clustering of risky behaviours [74]. This enables us to examine how different indicators of energy intake and expenditure are distributed, rather than using a simplified compound measure. Further, a data-driven approach does not require any assumptions of the existence or non existence of relationships between factors, and does not require a lengthy design phase via Delphi panels [185] or similar methods. Use of Bayesian classification models and other machine learning techniques in epidemiology is limited. This work seeks to provide a convincing argument of their utility and ability to model complex data.

The intended use of the population generated by the model is to aid policy makers by:

- Identifying population subgroups in which obesity related behaviours are clustered. These groups are likely to provide good targets for interventions.
- Aiding development of relevant targets to increase levels of physical activity and diet quality.
- Beginning to quantify the distribution of energy imbalance across a population. The size of the energy gap is unknown and controversial, but likely to be crucial in evaluating potential interventions.

More specifically, the model seeks to answer the following questions as worked examples to show the usefulness of this approach:

- How many people in Greater Manchester participate in no recreational physical activity? (with confidence intervals (CIs))
- How many participate in no walking, and typically consume less than one portion of fresh fruit or vegetables per day? (CIs)

6.3 Model Summary

The aim of this study is to create a model of health behaviours as determined by socio-demographic factors, to which detailed population specific socio-demographic data is applied to estimate the health behaviours in that population (see figure 6.1). The model is a Bayesian classifier, the structure and parameters of which are learned from HSE 2006 data. The applied socio-demographic data is from the 2001 UK Census. Here I apply the model to the Greater Manchester population, specifically the 10 Local Authorities that compose Greater Manchester; Bolton, Bury, Oldham, Rochdale, Salford, Stockport, Tameside, Trafford, Manchester and Wigan. Males and females are modelled separately.

A *Bayesian Classifier* is essentially a complete Bayesian Network [139]; both the topology H and the parameters θ are known. We can use our knowledge of the joint probability distribution specified to find the updated probabilities of a subset of variables when other variables are observed, subject to constraints of complexity. In this case, we wish to find the updated knowledge for the variables describing energy intake and expenditure; we denote this set $\mathbf{Y}$. The observed variables are the socio-demographic variables, denoted $\mathbf{X}$. The Health Surveys for England data contains variables for sets $\mathbf{X}$ and $\mathbf{Y}$, while the 2001 Census data contains only elements from set $\mathbf{X}$.

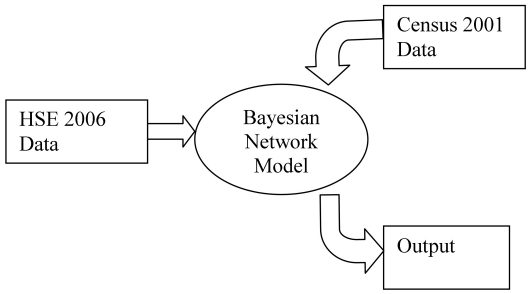


Figure 6.1: Schematic representation of how a Bayesian network model is used to estimate health behaviours in a sub-population

Construction of the Bayes classifier consists of two steps; first to derive a topology (H) for the network, then to assign parameters (θ). Methods are descibed in sections 6.5 and 6.6.

Participation bias is a known issue in the HSE; the individuals surveyed are not representative of any population. By conditioning across all socio-demographic variables in the model, the 2001 census data itself is used to express the relationships between the socio-demographic variables in set **X**. Consequently, any edges between nodes in the set **X** become irrelevant when the 2001 census data is applied.

6.4 Data

6.4.1 Overview

As stated above, this study uses data from the 2006 Health Surveys for England to learn the model, to which UK 2001 census data is applied. Both datasets are described in detail in Chapter 2. In brief, the UK census is a mandatory population survey conducted every 10 years to collect demographic information on the UK population, the most recent survey was conducted in 2001. The dataset used here is the Small Area Microdata (SAM), a 5% random sample of individual data. The Health Surveys for England are a series of annual surveys carried out to evaluate the health of the UK population. Participating households are selected at random from all over the UK. Households that decline to participate are not replaced; participa-

tion bias is therefore likely. HSE 2006 data is used as it focused on factors related to Coronary Heart Disease, and consequently collected data relevant to obesity.

The 2006 HSE data consisted of 21,399 individuals (10,007 males; 11,392 females). Following exclusion of individuals aged under 16 and over 75 (7,257), those failing to fill in self completion booklets (3,788), CORE 1 individuals over 65 (1,042), and those missing other variables (74), the final dataset contained 4,123 males and 5,115 females.

The 2001 SAM dataset contained 124,883 individual records within Greater Manchester (60,946 males; 63,937 females), representing a total population of approximately 1.78m. 34,969 individuals were excluded for being outside of the age range, and a further 851 missing other variables, leaving 45,004 males and 44,059 females.

6.4.2 List of Variables

The variables used from the HSE 2006 and 2001 Census are listed below. Variables in **X** were matched between the two datasets as closely as possible, and are shown in table 6.1. Variables in set **Y** are listed in table 6.2. Full descriptions of variables and derivations are provided in sections 2.1.3 and 2.2.3, asterisks (*) indicate that a variable has been edited specifically to match the 2001 Census data.

Socio-Demographic Variables		
HSE 2006 Variable	**Abbreviation**	**2001 Census Variable**
Sex	*Sex*	Sex
Age*	*Age*	Age
Dependent Children	*Dep.Cld*	Dependent Children
Marital Status	*Marital S*	Marital Status
Health Status	*Health*	Health Status
Social Status	*Social S*	Social Status
Economic Activity*	*Econ.S*	Economic Activity
Ethnicity	*Ethnic.*	Ethnicity
Education Level	*Educ.L*	Education Level

Table 6.1: List of 2001 Census and 2006 Health Surveys for England variables in this analysis; combing high and low resolution datasets

Energy Expenditure Variables	
Variable	**Abbreviation**
Recreational physical activity level	*Rec.PA*
Incidental physical activity level	*Inc.PA*
Occupational physical activity level	*Occ.PA*

Energy Intake Variables	
Variable	**Abbreviation**
Fried Food intake level	*FriedFd*
Cake/Sweets intake level	*Cakes*
Snack/Crisps etc. intake level	*Snacks*
Fruit and Vegetable intake level	*Frt Veg.*

Table 6.2: List of (unmatched) 2006 HSE variables used in this analysis

6.5 Obtaining the Classifier Structure

6.5.1 Methods

To avoid unmanageable complexity, it is preferable to use a single topology H for the Bayesian classifier as opposed to averaging over several network structures. One possible approach is to derive a topology H by maximising some scoring criterion. However, this does not consider the entropy of network structures, *i.e.* a large number of equivalent DAGs may be excluded in favour of a single DAG with a higher scoring criterion despite their combined probability being greater.

Exhaustive evaluation of the space of all possible networks G is unfeasible; a Metropolis Hastings sampler is used to traverse and sample the space. Topologies are grouped according to equivalence classes [152]. An equivalence class is the set of all DAGs that degenerate to the same *Partially Directed Acyclic Graph* (PDAG). All arcs that do not participate in a 'V-structure' before or following reversal are considered reversible, and represented by an undirected edge in a PDAG. Thus a PDAG is a mixture of directed and undirected arcs. It is difficult to evaluate the probability of a PDAG directly, consequently this numerical method is applied.

The DAG topology to be used in the Bayes classifier is constructed from the most commonly sampled, and therefore the most probable, PDAG. All undirected arcs between $\mathbf{X}$ and $\mathbf{Y}$ are orientated so that members of $\mathbf{X}$ are parents, arcs between members of $\mathbf{Y}$ are directed according to the algorithm provided by Chickering [152].

In order to reflect the network's role as a classifier of the set $\mathbf{Y}$ rather than a

model of the joint distribution, a specialised network topology scoring criterion is used. Heckerman describes the use of the probability of the data given a model topology H as a measure of model fit [142]. This is specified in section 3.3.1 and provided below:

$$\Pr(D|H) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}. \tag{6.1}$$

Essentially, this is the product of marginal likelihoods over possible θ for all node (i) and input (j) combinations. This marginal likelihood given H is combined with a prior over network structures (eq. 3.13) to provide the scoring criterion. Modified scoring criterions for model selection are discussed in Heckerman [142]. The adjusted criterion ignores the likelihood contribution of the members of $\mathbf{X}$, the network is scored on its ability to predict the states of the set $\mathbf{Y}$ only:

$$\Pr(D_Y|H) = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \right\}^{\mathbb{I}_{\mathbf{Y}}(i)}. \tag{6.2}$$

where

$$\mathbb{I}_{\mathbf{Y}}(i) = \begin{cases} 1 & \text{if} \quad i \in \mathbf{Y} \\ 0 & \text{if} \quad \text{otherwise.} \end{cases}$$

Advantages of this modified criterion include:

- Ensures that arcs that explain correlations between socio-demographic variables do not contribute to the evidence score of the topology.
- Preferentially orientates arcs from set $\mathbf{X}$ to $\mathbf{Y}$.
- Flattens the probability distribution, resulting in easier mixing and convergence.

In addition to the modified scoring criterion the prior over network structures is also modified from that described in eq. 3.13. Due to entropy, 'quiet' edges become prevalent in the network structures if not penalised. Quiet edges do not influence the marginal likelihood component of the scoring criterion, although the prior over network structures is influenced. However, this influence of the complexity penalising prior is insufficient to prevent their occurrence. These edges are problematic as they have negative effect on mixing by restricting the moves available

to the sampler. A penalty on edges between members of $\mathbf{X}$ is imposed; edges incident to members of the set $\mathbf{X}$ incur a log penalty of $-1,000$, effectively reducing the probability of the network by a factor of $e^{1,000}$. This practically eliminates the prevalence of 'quiet' edges. This penalty value was chosen following consideration of the range of the scoring criterion. As the marginal likelihood is directly proportional to the number of observations, this penalty value should be scaled accordingly if this methodology is applied to other datasets.

Using the adjusted scoring criterion, a Metropolis Hastings sampler was implemented over the space of network structures. Males and females were analysed separately. Four runs were performed for both male and female data. Two runs were seeded from an empty DAG, and two from the optimal DAG as derived from simulated annealing optimisation. A burn-in period of 5×10^5 iterations was performed before a sampling period of 5×10^5 iterations with samples every $1,000$. Providing a sample of 500 DAGs for each run.

The move library consisted of the 5 moves described in section 4.4.1, with the following frequencies:

- Add Arc: 0.4.
- Delete Arc: 0.4.
- Grzegorczyk-Husmeier REV: 0.1.
- Switch Arc: 0.05.
- Multiple Reversal: 0.05.

6.5.2 Results of the Metropolis Hastings Sampler

Graphs of the evidence traces and edge relation features associated with the samplers can be seen in the appendix. In both the male and female data, one of the empty initialisations failed to converge (appendices D.1 and D.3). There was also an issue with one of the informed initialisations becoming stuck in a sub-optimal region in the female data. Despite these mixing issues, there was a clearly identifiable preferred PDAG for both males and females. In both cases the preferred PDAG was consistent with the DAG derived by simulated annealing optimisation. Table 6.3 displays the counts of the different PDAGs sampled (from chains that successfully converged). Each variable was assigned an index from 0 to 14, with the order: 0 DepChld, 1 MaritalS, 2 HealthS, 3 Ethnicity, 4 EducL, 5 Age, 6 EconA, 7

Social S, 8 Snacks, 9 Cakes, 10 FriedFd, 11 FrtVeg, 12 Rec PA, 13 Inc PA, 14 Occ PA.

Following translation of the selected PDAG into a DAG (as specified above) network topologies for the Bayes classifiers are displayed in figures 6.2 and 6.3. Nodes are colour coded, with those representing socio-demographic variables in blue, and those representing behavioural indicators in yellow. Variables not included in the final model are pictured for completeness.

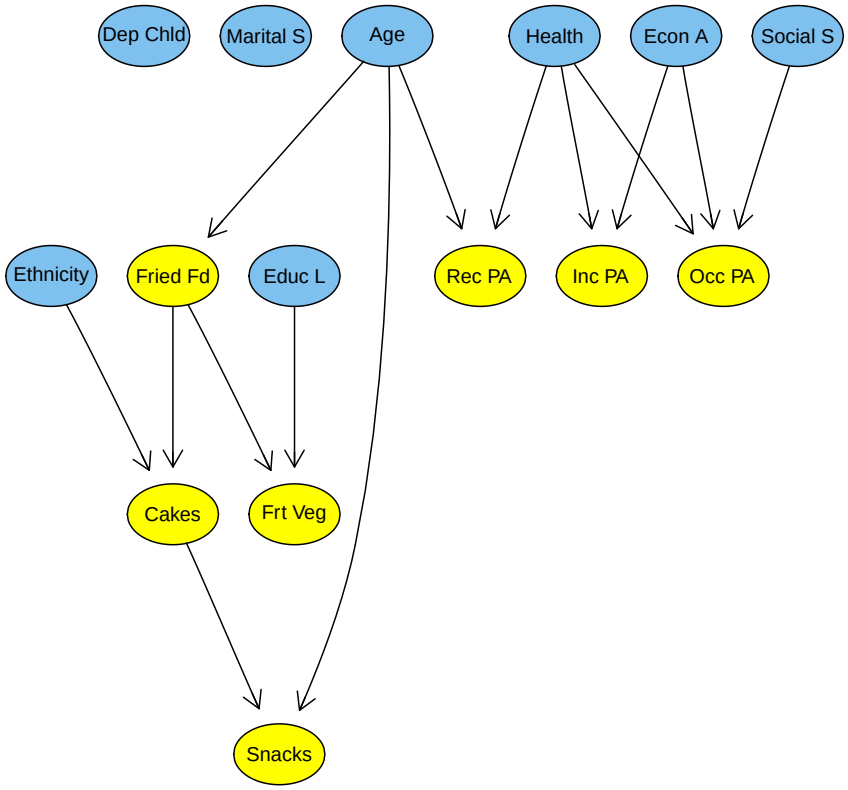


Figure 6.2: Topology of the Bayesian health behaviour classifier discovered following Metropolis Hastings sampling (males). Blue nodes denote socio-demographic variables; yellow nodes, behavioural indicators

(a) Males

PDAG	frequency
2-12, 2-13, 2-14, 3-9, 4-11, 5-8, 5-10, 5-12, 6-13, 6-14, 7-14, 9-8, 10-9, 10-11	1209
2-12, 2-13, 2-14, 3-9, 5-8, 5-10, 5-12, 6-13, 6-14, 7-14, 9-8, 10-9, 10-11	160
2-12, 2-13, 2-14, 3-9, 5-8, 5-10, 5-12, 6-13, 6-14, 7-11, 7-14, 9-8, 10-9, 10-11	79
2-12, 2-13, 2-14, 3-9, 4-11, 5-8, 5-10, 5-12, 6-13, 6-14, 7-14, 9-8, 10-9	30
2-12, 2-13, 2-14, 3-9, 3-13, 4-11, 5-8, 5-10, 5-12, 6-13, 6-14, 7-14, 9-8, 10-9, 10-11	11
2-12, 2-13, 2-14, 3-9, 3-11, 5-8, 5-10, 5-12, 6-13, 6-14, 7-14, 9-8, 10-9, 10-11	5
2-12, 2-13, 2-14, 3-8, 3-9, 4-11, 5-8, 5-10, 5-12, 6-13, 6-14, 7-14, 9-8, 10-9, 10-11	5
0-10, 2-12, 2-13, 2-14, 3-9, 4-11, 5-8, 5-10, 5-12, 6-13, 6-14, 7-14, 9-8, 10-9, 10-11	1
2-11, 2-12, 2-13, 2-14, 3-9, 5-8, 5-10, 5-12, 6-13, 6-14, 7-14, 9-8, 10-9, 10-11	1
2-12, 2-13, 2-14, 3-8, 3-9, 3-13, 4-11, 5-8, 5-10, 5-12, 6-13, 6-14, 7-14, 9-8, 10-9	1
2-12, 2-13, 2-14, 3-9, 3-13, 5-8, 5-10, 5-12, 6-13, 6-14, 7-14, 9-8, 10-9, 10-11	1

(b) Females

PDAG	frequency
2-12, 2-13, 2-14, 3-10, 4-11, 4-12, 5-8, 5-11, 6-14, 9-8, 10-9, 11-10, 14-13	1246
2-13, 2-14, 3-10, 4-11, 4-12, 5-8, 5-11, 6-12, 6-14, 9-8, 10-9, 11-10, 14-13	164
2-13, 2-14, 3-10, 3-12, 4-11, 4-12, 5-8, 5-11, 6-12, 6-14, 9-8, 10-9, 11-10, 14-13	81
2-12, 2-13, 2-14, 3-9, 3-10, 4-11, 4-12, 5-8, 5-11, 6-14, 9-8, 10-9, 11-10, 14-13	10
2-13, 2-14, 3-9, 3-10, 4-11, 4-12, 5-8, 5-11, 6-12, 6-14, 9-8, 10-9, 11-10, 14-13	2

Table 6.3: Observed PDAG frequencies of Bayesian network topologies of data from the 2006 HSE over the Metropolis Hastings sampling process. PDAGs are represented by a list of arcs using the node IDs provided in the previous section

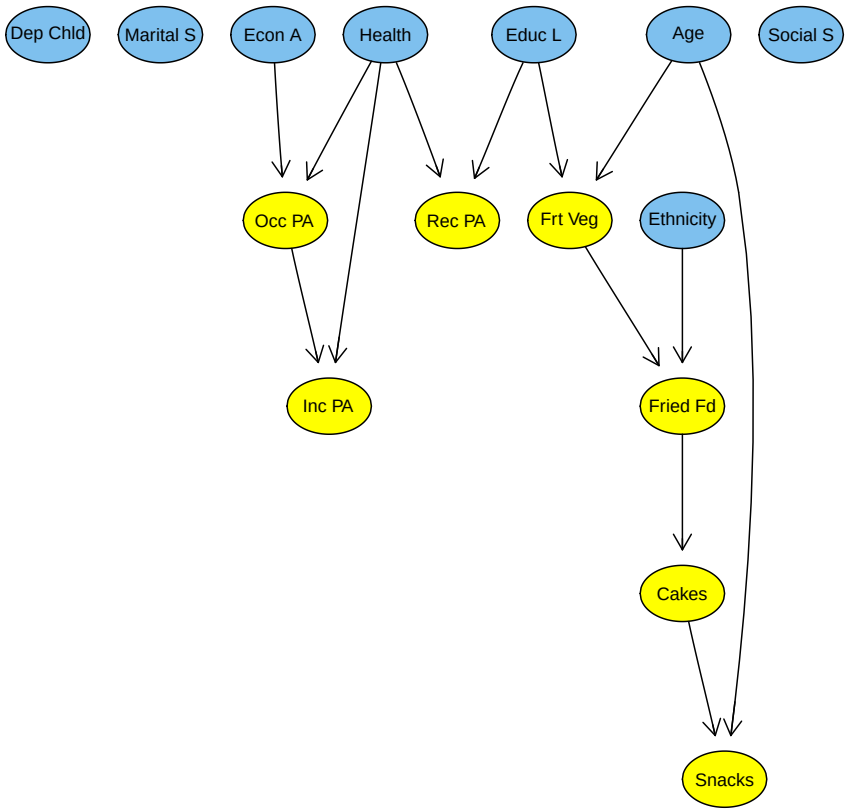


Figure 6.3: Topology of the Bayesian health behaviour classifier discovered following Metropolis Hastings sampling (females). Blue nodes denote socio-demographic variables; yellow, behavioural indicators

6.6 Assigning Model Parameters

Given a network structure, the next step is to assign parameters to the model. Learning of Bayesian Network parameters is discussed in section 3.3.2. In the multinomial case the parameter vector for each node-input combination (θ_{ij}) follows a Dirichlet distribution. It is straightforward to identify each value of θ_{ij} that maximises the posterior probability, however the associated uncertainty is also of interest. This point may be illustrated by considering two Beta-distributions (the simple binomial case of the Dirichlet) in figure 6.4: both Beta distributions share the same maximum likelihood value of 0.5. However in the case of Beta(10,10) we can be more confident of its value than Beta(2,2). The selection of the maximum likelihood value results in the loss of a substantial amount of information, increasing the risk of bias due to small group effects.

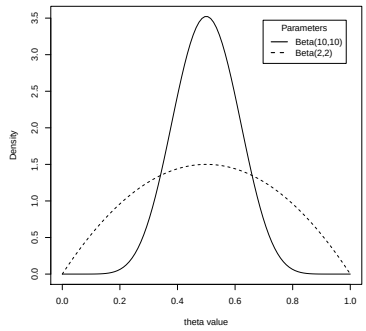


Figure 6.4: Comparison of the shape of two Beta-distributions

The posterior distribution of θ_{ij} can be specified exactly by a Dirichlet distribution Dir($n_1 + \alpha_1, n_2 + \alpha_2, ..., n_r + \alpha_r$) with parameters from HSE counts and designated pseudocounts. The selection of the distribution of θ_{ij} rather than a point estimate allows the calculation of confidence intervals associated with this uncertainty.

6.7 Application

Following sections 6.5 and 6.6, the model is complete with topology and parameters. The next step is the application of the 2001 Census data to answer specific queries. The topology of the model is fixed, and not query dependent. The 2001

Age	Males	Females
16-19	65,030	63,105
20-29	159,716	165,487
30-39	191,230	196,436
40-49	160,161	161,709
50-64	210,083	212,612
65-74	90,880	105,405
Total	877,100	904,754

Table 6.4: Counts of individuals by age and sex in the 2001 Census (Greater Manchester)

Census Small Area Microdata dataset consists of 89,063 individuals between the ages of 16 and 74, drawn randomly from a population of 1.78 million in Greater Manchester (table 6.4). An estimated 96% of the population were recorded by the Census [135], representing the most complete sample of socio-demographic characteristics available. The method is described below, and also expressed as pseudocode in figure 6.5 for clarity.

A query is submitted to the model, containing the nodes for which the distribution in the true population is of interest. The query may also restrict the estimate to specific subgroups. Given a query, the relevant socio-demographic nodes are identified, *i.e.* the parents of the nodes of interest to the user. Every possible combination of the states of the relevant nodes provides us with a socio-demographic group, depending on the number of nodes of interest the number of these groups may be significant. These groups are termed *metagroups*. Within each metagroup, the distributions of the nodes of interest are independent, as the states of the parents are fixed in each case. Where a node of interest has a member of $\mathbf{Y}$ as a parent, this requires an imputation step; this is described later. This process involves the use of several Dirichlet distributions, in each case the pseudocount (α) is set at 1.0.

The Census data available is a 5% sample of the overall data, consequently the numbers of the true population in each metagroup are unknown. A Bayesian approach to this problem is taken; counts in each metagroup from the 2001 SAM Census data ($n_1^c, ...n_z^c$) are used to inform a Dirichlet distribution as to the true probabilities of a randomly drawn individual from the Greater Manchester population being a member of each metagroup, where z is the number of metagroups. A

Query: Identify Nodes of Interest.

Indentify parents of these nodes. Each combination of parent outcomes forms a metagroup.

If parents are not members of set $\boldsymbol{X}$, an imputation step is necessary.

For each sample {

Perform imputation (if required).

Use counts in each metagroup from census data (and imputation if required) to inform a Dirichlet distribution representing probability of each individual in true population being in each metagroup.

Draw a sample from this Dirichlet. This sample is used to inform N samples from a multinomial distribution, where N is the number of individuals in the true population. The resulting counts estimate the numbers within each metagroup.

For each metagroup {

Determine the Dirichlet distribution for the outcome of each node of interest from HSE data. Sample from each of these distributions and take the product as an estimate of the joint distribution of the nodes of interest.

Use this product to inform a multinomial distribution, from which we draw a number of samples according to our previous estimate on numbers within metagroups.

The results of these samples are stored.

}

Results are summed over all metagroups.

}

Figure 6.5: Summary of Dataset Combination Method in Pseudocode

random sample is then drawn from this Dirichlet distribution to inform N samples drawn from a multinomial distribution, where N is the total number of individuals in the population (877,100 males and 904,754 females). This generates an estimate of the number of people ($\hat{n_1}....\hat{n_z}$) in each metagroup in the true population.

Next, each of these simulated individuals is assigned to a category of each node of interest dependent on their metagroup. This is also done by sampling from a multinomial distribution- this time informed by a sample from a Dirichlet distribution defined by the HSE data. For each metagroup, the relevant Dirichlet parameters from the HSE are determined, a sample from which is drawn and combined with that from the other nodes of interest to produce a joint distribution over all combinations of queried nodes; recall these nodes are independent if all parent nodes are instantiated. This informs a multinomial distribution of $\hat{n_\iota}$ individuals, providing a single estimate of the number of individuals possessing each possible comnination of the node(s) of interest.

This process is repeated until 10,000 estimates are drawn- allowing approximation of a mean and confidence intervals. A simulation approach is necessary, as analytical evaluation of this two-tiered Bayesian problem is extremely complex. The method described here is implemented in *R*, an open source statistical package [186]. The R script is included in appendix C. The time taken to draw 10,000 samples depends heavily on the number of metagroups in the analysis, which is largely determined by the number of nodes of interest. The number of metagroups can be reduced by restricting the query to population subgroups. The simulation in question 1 was generated in approximately 12 seconds on a dual core 2.4GHz 2GB RAM machine, while question 2 took in the order of 30 minutes.

As mentioned above, an imputation step is required where a node of interest has a non socio-demographic node as a parent. This is because the numbers in each metagroup cannot be estimated, as the parent is not present in the Census data. The imputation step proceeds as follows; the ultimate socio-demographic parents of the missing node are identified. These may not be immediate parents as numerous indicator variables may be descendants of one another (as in the female classifier), in this case, several imputation steps may be required. For each group of the combination of these parent variables, the HSE data is used to derive the relevant parameters of the Dirichlet distribution of the outcome probabilities of the node. A sample from this Dirichlet is used to inform a multinomial that assigns each individual in this group a value for the non socio-demographic node. This sample is then used exactly as if it were present in the Census dataset. Compared to

the above process this imputation step is computationally inexpensive; it is feasible to perform a separate imputation for each sample estimate.

Question 1

How many people in Greater Manchester participate in no recreational physical activity? Or, in the context of this study, how many people in Greater Manchester would fall into the lowest category (*i.e.* none) of the recreational physical activity (RPA) variable?

The parents of the nodes of interest are health and age in males, and health and education level in females. All of these parents are members of the set $\mathbf{X}$ and thus no imputation step is necessary. The RPA node has 18 input levels in the male classifier and 12 in the female. As there is only one node of interest (RPA), these are also the number of metagroups. Following the process described in the previous section- $10,000$ estimates were generated for the distribution of recreational physical activity behaviour of the Greater Manchester population. Of the $877,100$ males; a mean of $468,045.7$ (53.4%) fell into the lowest RPA category, $CI_{95\%} : \{454,506; 481,477\}$. In females a mean of $559,635.5$ of the total $904,754$ (61.9%) fell into the lowest RPA group $CI_{95\%} : \{547,892; 571,343\}$. Of the total population ($1,781,854$ individuals), a mean of $1,027,681$ (57.7%) was calculated $CI_{95\%} : \{1,009,852; 1,045,487\}$.

Question 2

How many individuals do little or no walking, and fail to eat at least one portion of fruit or vegetables a day? Or, how many people in Greater Manchester fall into the lowest incidental physical activity category, AND fall into the lowest consumption group for fruit and vegetables.

This is a much more complex question than the previous one, requiring an imputation step and exhibiting a large number of metagroups. In the male classifier, *Fried fd* is a parent of *Frt Veg*, in females *Occ PA* is a parent of *Inc PA* cases. Both will require an imputation step. Further, the number of metagroups is large, 288 in females, and 192 in males. Due to this complexity the simulation took approximately 30 minutes to complete. In females, an average of $21,694.9$ of the $904,754$ (2.4%) fell into both categories of interest ($CI_{95\%} : \{20783.0; 22626.0\}$). In males the average number of individuals in the lowest categories was $28,460.6$ (3.2%), with confidence interval $CI_{95\%} : \{26,851.0; 30,098.1\}$. In total an average

of 50, 155.5 or 2.8% of individuals fell into the lowest categories for both incidental physical activity and fruit and vegetable consumption.

6.8 Discussion

This study exploits various properties of Bayesian networks to estimate the obesity related behaviour of a sub population in ways that may be more robust than alternative linear methods. Data from health surveys is subject to participation bias, which limits its applicability to real populations. This approach, while not circumventing participation bias, projects the results from a national survey onto a real sub-population. The uncertainty that occurs when a group is highly under-represented in the original survey, is reflected by uncertainty in the behaviour of that group when the model is applied to the true population. This is a significant advantage of this method. For example in the HSE, individuals with no educational qualifications made up 22.7% of the sample, compared to 31.2% in the Census data, consequently, the health behaviour of this group is associated with higher uncertainty than the over-represented higher education group. This work may be relevant to Local Authorities to identify the health needs of their population. The health needs of regions are likely to vary, with population demographics likely to play an important role in what interventions are appropriate.

Beyond population health, the work may have wider applications. There are numerous situations where the results from a detailed survey on a small scale may be applied to a larger lower-resolution dataset, Molitor *et al* combined several regression models using a Bayesian network to predict wider scale implications of water quality on foetal development [187]. The flexibility of Bayesian networks makes them well suited to these types of analyses.

The inclusion of more variables would have improved the predictive power of the model. However the variables were limited to those that could be successfully matched between census data and the HSE. Throughout this study I have assumed that the matched variables are directly equivalent, this may not hold. Although most are based on well defined criteria, such as age, sex, ethnicity, and economic status, others such as self reported health or dependent children are less clearly defined. Table 2.1 compares the questions from which these variables are derived. This may contribute to bias in some groups.

The approach described here is entirely data driven, meaning no investigator input was involved in the design of the model. The method implemented was just

one of a number of possible methods of constructing a classifier. A 'Naïve Bayes classifier' would have been simpler to generate, but the inclusion of the parameter distribution in the model is a major strength, enabling the quantification of uncertainty. The maximisation of the scoring criterion is simpler approach than the derivation of the most probable PDAG, and when tested yielded the same result (data not shown). An alternative knowledge-based approach could have employed a domain expert to design the network structure, but this would have been susceptible to expert bias and is resource intensive.

Despite the modified scoring criterion, mixing and convergence over DAG structures was not straightforward. The heavy penalties imposed on arcs between members of $\mathbf{X}$ may be a contributory factor. When a REV move is performed in the MCMC process, the local scores that determine the new parentset (see section 4.4.2) are not penalised. As a result, many attempts are made to impose arcs between members of the set, which are rejected due the penalty on the topology prior. This reduces the effectiveness of the REV move enormously. The additional mixing complexity associated with the modified criterion may encourage a more straightforward approach, such as maximising the scoring criterion as described above. The simulation process is computationally intensive, especially for complex queries. The R interpreter is less efficient than programming languages such as C#, much of the processing time is spent manipulating matrices and generating counts from data; the use of alternative technologies may improve performance considerably.

The model is capable of estimating the behaviour of any group over any variable or combination of variables. Further development could involve the presentation of this approach as a tool, requiring the installation of a user interface that will allow the model to be queried directly by users.

Chapter 7

Identification of Predictors of Waist to Hip Ratio in UK Adults using Bayesian Networks

7.1 Overview

In this chapter I use Bayesian model averaging (BMA) to investigate conditional dependencies between variables representing body fat distribution and its putative determinants. The dataset is derived from the 2006 Health Surveys for England (HSE). Particular attention is paid to deterministic relationships with waist to hip ratio (WHR). Analysis using BMA is compared with the more standard epidemiological technique of generalized linear models. The standard regression technique proves to be more sensitive in identifying determinants, but BMA provides a more complete explanation of relationships present in the data as a whole.

7.2 Background

Obesity is the carriage of excess adipose tissue which impairs the body's ability to function correctly. This manifests as increased risk of a number of common chronic diseases including Type II diabetes, coronary heart disease (CHD) and various cancers. Obesity is most commonly defined in terms of weight adjusted for height; body mass index (BMI). When risk associated with excess adiposity is reported, BMI is the accepted measure of overweight and obesity [1]. However BMI is a flawed indicator of obesity related disease risk for individuals [188]. Excess visceral fat, adipose tissue that is located in the abdominal cavity around the vital organs, is associated with elevated risk of diabetes and CHD *independently* of BMI [11, 189–191]. Waist hip ratio (WHR) provides a useful indicator of the ratio of visceral fat to body size. The fundamental driving force behind fat deposition is clear; an imbalance between energy consumed and energy expended. However, less clear are the determinants of *where* body fat is deposited. There is insufficient data from longitudinal studies to reveal the determinants of fat distribution in full. The most ready source of relevant data is large cross sectional studies such as the Health Surveys for England. Aside from age, sex [192, 193] and ethnicity [194–196], attempts to uncover determinants of fat deposition have had little consistent success. Smoking has been identified by several studies as associated with higher WHR adjusted for BMI [197–200], however no putative mechanism has been proposed [201]. The issue of smoking associations are complicated by a tendency for ex-smokers to gain weight after cessation of smoking [202]. Other factors that have been put forward as potential contributors to body fat deposition are alcohol intake [197, 198, 203], menopause [192, 204], and exercise and diet

composition [205].

This study has two main aims:

- To use Bayesian model averaging to identify determinants of body fat distribution.
- To compare this method with a Generalized Linear Modelling approach in this context.

This chapter is organised as follows; section 7.3 describes the methodology implemented, and the results of these analyses are presented in section 7.4. A discussion is provided in section 7.5.

7.3 Approach and Methods

7.3.1 Overview

In this chapter I sample from the posterior distribution of Bayesian network topologies given data to identify features shared by the most likely network structures. This is known as Bayesian model averaging (BMA), and is detailed in section 3.4. These structural features provide an estimate of the conditional dependencies present in the data. A Metropolis Hastings sampler is used to traverse and sample from the space of possible networks (section 3.5).

In a Bayesian network, conditional dependencies are encoded by directed arcs between nodes; the structure of the network is essentially a set of assertions of conditional dependence between variables. A useful feature of Bayesian networks is the ability to distinguish between correlation and conditional dependence [124]. This study uses the presence of arcs to identify conditional dependencies present in a dataset, specifically for factors that influence WHR once the large explanatory effect of BMI is adjusted for. The results from this analysis are *edge relation features* (ERFs), an estimate of the posterior probability of each arc. A high ERF indicates that the arc is important in explaining the observed data. Bayesian topology space approach are compared with the results from a more usual epidemiological tool; generalized linear models [206] (GLM). This comparison will enable us to evaluate the utility of the Bayesian topology approach in answering a real question within a typical epidemiological context.

Male and female data are analysed separately owing to likely different determinants of fat deposition [192, 193] and lack of comparability between WHR values.

Further, due to a strong evidence base suggesting different body fat distribution by ethnicity [194–196] only individuals self identified as of white ethnicity are included.

7.3.2 Data

The data source is the 2006 Health Survey for England, full details of which are provided in Chapter 2. Variables were selected following a literature review of studies examining potential determinants of fat deposition. The variables used in this analysis, as described in section 2.1.3 are listed in table 7.1. The 2006

Socio-Demographic Variables	
Variable	**Abbreviation**
Sex	*Sex*
Age (groups)	*Age*
National Statistics Scoio-Economic Classification	*NS-SEC*
Economic status	*Econ.S*
Education level	*Educ.L*
Energy Expenditure Variables	
Variable	**Abbreviation**
Recreational physical activity level	*Rec.PA*
Incidental physical activity level	*Inc.PA*
Occupational physical activity level	*Occ.PA*
Energy Intake Variables	
Variable	**Abbreviation**
Fried food intake level	*FriedFd*
Cake/sweets intake level	*Cakes*
Snack/crisps etc. intake level	*Snacks*
Fruit and vegetable intake level	*Frt Veg.*
Physical Variables	
Variable	**Abbreviation**
BMI (groups)	*BMI*
WHR (groups)	*WHR*
Period status	*PeriodS*

Table 7.1: List of variables used in this analysis; using Bayesian networks to identify factors that influence body fat distribution

HSE surveyed 21, 399 individuals (10, 007 males; 11, 392 females). The following

Age Group	Males	Females
16-24	230	331
25-34	558	459
35-44	908	748
45-54	787	669
55-64	858	760
60-74	321	314
Total	3,662	3,281

Table 7.2: Counts of individuals in 2006 Health Survey for England by age category

individuals were excluded from this analysis; those aged under 15 or over 74 years (4,284; 4,475), non-white individuals (572; 700), CORE 1 individuals aged over 65 (see section 2.1 for details) (384; 409), individuals failing to fill in the self completion (SC) booklet (1,200; 1,442), individuals who failed to provide reliable weight information (266; 437), and those missing other variables present in the model (20; 267). In total 6,943 individuals were available for analysis; 3,281 males and 3,662 females. Counts in each age category are provided in table 7.2.

In order to meet the assumptions of a GLM, and to ensure a fair comparison of the two methods, some variables were implemented in the GLM as continuous. These were Age (years), BMI, and WHR. All other variables were treated as categorical.

7.3.3 Implementation of the Metropolis Hastings Sampler

A Metropolis Hastings sampler was implemented for both males and females over the datasets described in the previous section. Following a burn-in period of 5×10^5 iterations, a sampling period also of 5×10^5 iterations was performed, with samples taken every 1,000 iterations. Network topologies were scored using the scoring criterion described in section 3.3.1, with pseudocounts set at 1.0. The move library was as described in section 4.4.1 with specified probabilities:

- Add Arc: 0.4.
- Remove Arc: 0.4.
- Grzegorczyk-Husmeier REV move: 0.1.
- Switch Arc: 0.05.

- Multiple Reversal move: 0.05.

For each dataset, four runs of the sampler were performed, two chains were initialised from an empty network, and two from the optimal network as derived using simulated annealing. This allows monitoring of mixing and convergence of the chains, which is known to be problematic in many cases.

7.4 Experimental Results

Results are presented graphically; an undirected edge between two nodes indicates that an edge between the two nodes was observed in the sample. The ERF is stated where it is not equal to 1.0, which is denoted by a black arc. Edges are colour coded by ERF value for clarity: Black, ERF = 1.0; Red, ERF = $0.1-0.99$; Yellow, ERF = < 0.1. In addition to edge relation features, optimal topologies are also included in figures 7.3 and 7.4. Log Evidence traces and scatter plots of edge relation features between initialisations are provided in appendix E.

7.4.1 Metropolis Hastings Sampling

In the male data one of the empty initialisations failed to converge (E.1). Apart from this one run, estimates of edge relation features were similar (E.2). The female data displayed better mixing, with both empty initialisations converging (E.3). However, chains experienced short periods of becoming stuck in sub-optimal regions. This explains the slight discrepancy in the observed edge relation features between the two chains (E.4).

As illustrated by figures 7.1 and 7.2, the derived edge relation features indicate *Age* and *BMI* as the major explanatory factors of *WHR*. In all topologies sampled from male and female data an edge was observed between *Age* and *WHR*, and *BMI* and *WHR*. *Age* was highly connected across all networks to many other variables. The presence of the *Age* to *WHR* edge indicates that age is a better predictor of WHR than it is of BMI, and suggesting conditional independence of BMI and age. The observed relationship between age and WHR may be explained by life-course physiology and/or behaviours. The results suggest that as individuals age they are prone to a higher WHR, even within BMI groups. BMI conflates the loss of lean mass as part of aging with changes in fat mass over the life course. In the absence of measures of lean and fat mass, conditioning on WHR goes some way to disaggregating these relations.

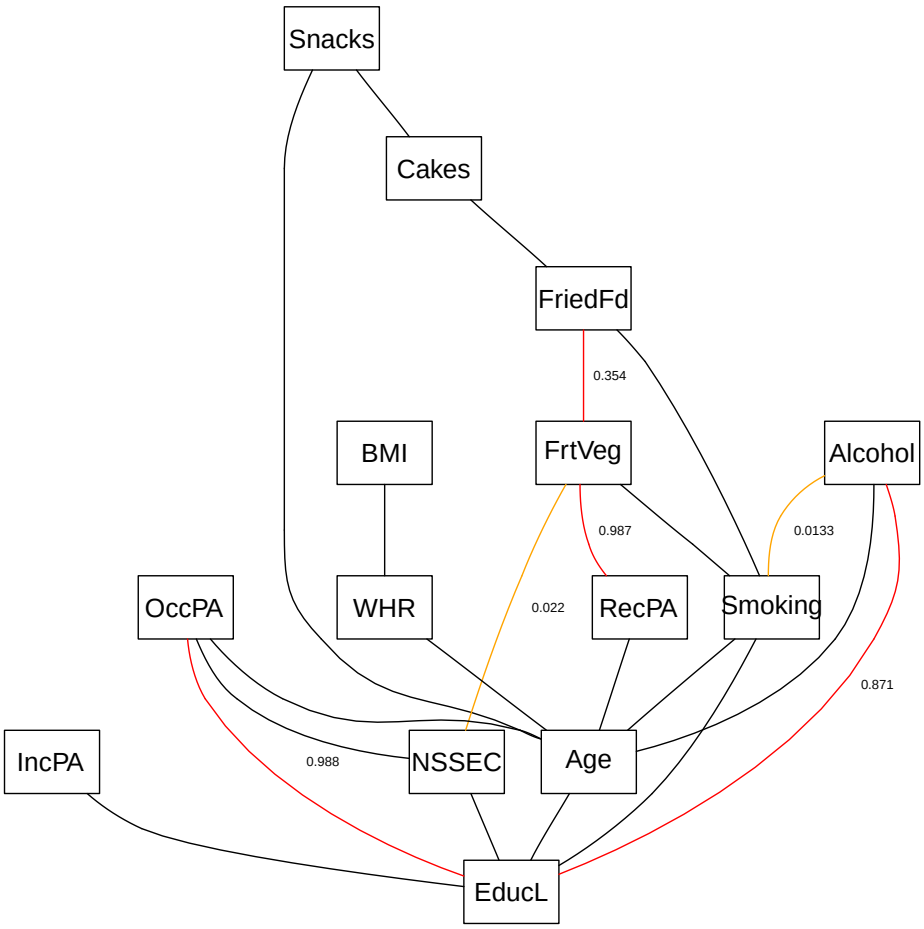


Figure 7.1: Relationships between fat distribution and eating, physical activity and socio-demographic factors in males presented as the average Bayesian network topology, derived from Metropolis Hastings sampling (2006 HSE data)

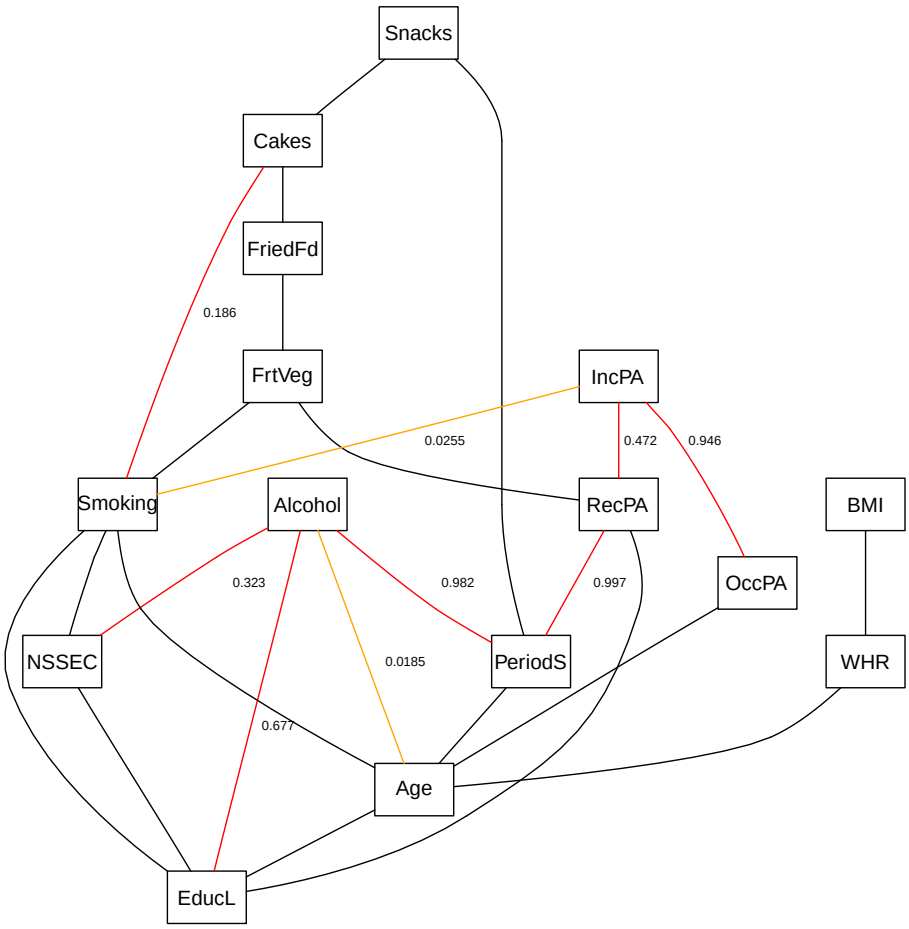


Figure 7.2: Relationships between fat distribution and eating, physical activity and socio-demographic factors in females presented as the average Bayesian network topology, derived from Metropolis Hastings sampling (2006 HSE data)

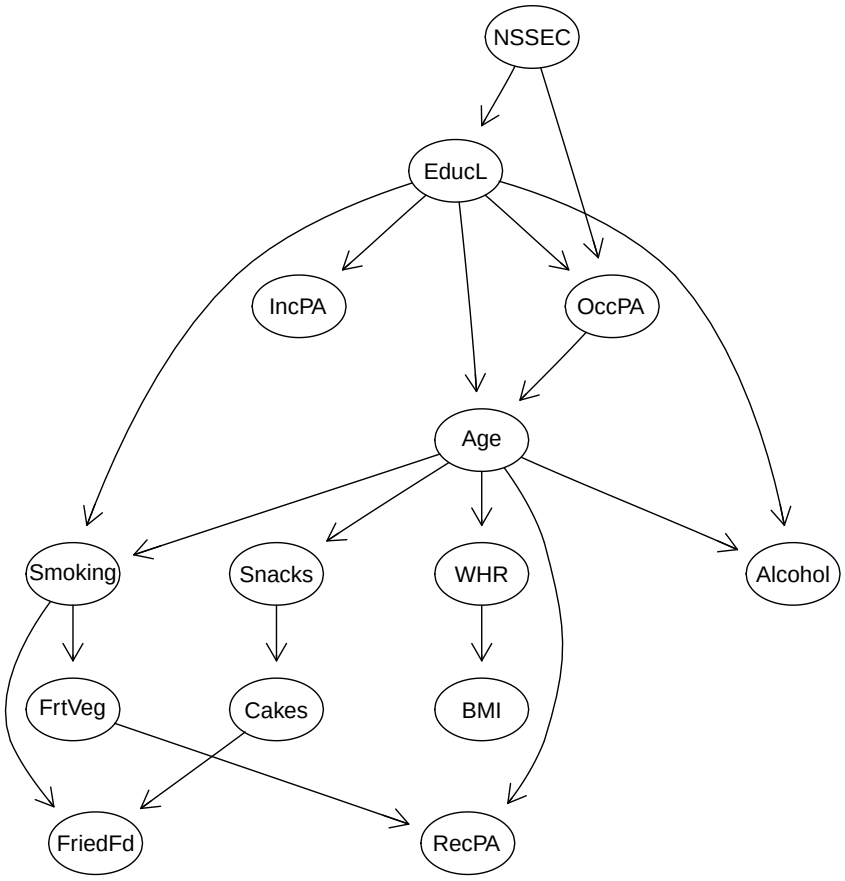


Figure 7.3: Optimal Bayesian network topology of obesity related factors from 2006 Health Surveys for England data discovered using simulated annealing (Males)

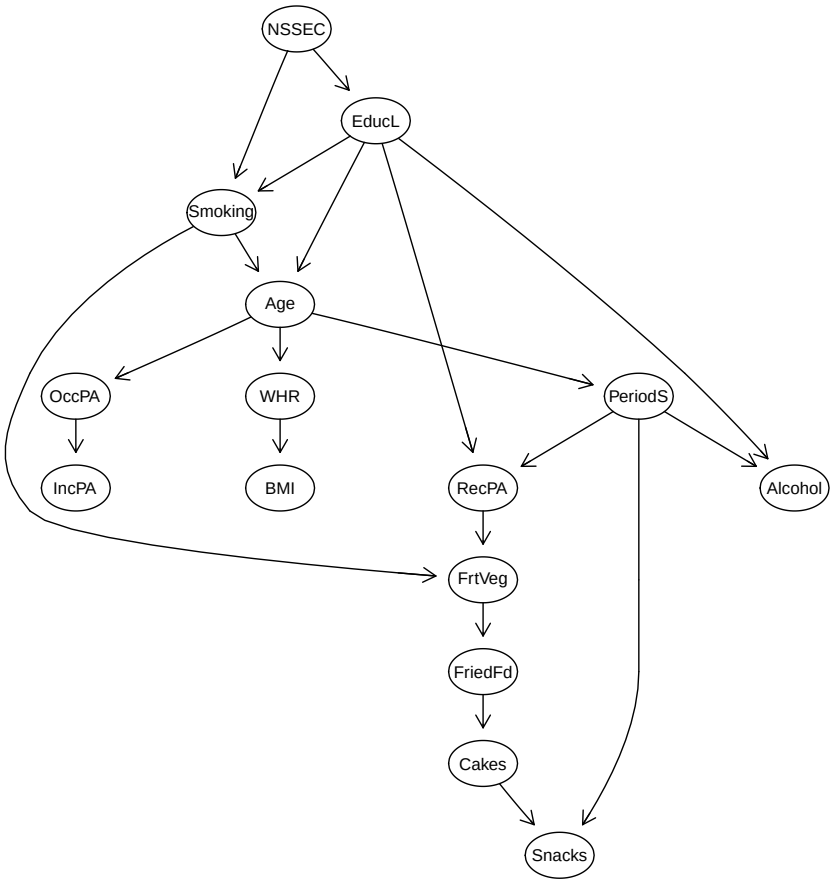


Figure 7.4: Optimal Bayesian network topology of obesity related factors from 2006 Health Surveys for England data discovered using simulated annealing (Females)

Notably, all other variables are conditionally independent of *WHR* and *BMI* given age in all sampled network topologies. Age is strongly correlated the other socio demographic variables, which in turn associate strongly with behavioural variables. Correlation of *Age* and *BMI* with *WHR* may dominate weaker dependency relations present.

Between males and females, similar network dynamics are observed. The *Period status* variable is highly connected, this is probably due to a proxy effect of its extremely close relationship with age. Females also show a correlation of alcohol use with education and social class, which is not observed in males.

7.4.2 Application of a Generalized Linear Model

As discussed in the above section, variables were not directly equivalent for the GLM analysis, several variables were treated as continuous as discrete categorisation of variables is a limitation of the Bayesian topology method.

WHR was defined as the dependent variable. All variables were included in the model, and backwards elimination stepwise regression carried out; variables were removed in order of lowest $\boldsymbol{z}$ value until only those significant at $\alpha = 0.05$ remained. This stepwise approach has numerous limitations [207], but in order to ensure a fair comparison of approaches it must be data driven. For males, the final model contained the variables, BMI, Age, Smoking, Education, Cake and Fruit and Vegetable intake. An interaction between Age and Smoking was also successfully fitted. Relevant $\boldsymbol{z}$ scores and significance statistics are included in table 7.3. The GLM identifies age and BMI as the major predictors of WHR- in agreement with the Bayesian topology model. However, several other significant indicators were also highlighted; current smokers and ex-regular smokers displaying a slightly higher WHR than never smokers. Increased levels of recreational physical activity have a dose response effect resulting in lower WHR for those that do more recreational exercise. Fruit and vegetable intake results in a higher WHR with lower consumption. Cake intake also shows a higher WHR with lower intake, while fried food intake has the opposite effect. Interestingly, the socio demographic factor of education level also appears to have an influence; this may be due to a correlation with behaviour that is not incorporated in the model, those with no qualifications have a slightly higher WHR as compared to those with higher education.

In females the final model contained Age, BMI, Smoking, Education level, NSSEC, and Snack and Fried food intake. Again BMI and age were the most sig-

	Coef.	Std. Err.	*z*	P> *z*	95% Conf. Interval
Cakes (vs ≤1pw)	-	-	-	-	-
(1-2pw)	-0.00524	0.001997	-2.63	0.01	(-0.0092,-0.0013)
(3-5pw)	-0.00525	0.002337	-2.25	0.03	(-0.0098,-0.0007)
(5+pw)	-0.00819	0.003854	-2.13	0.03	(-0.0157,-0.0006)
Fried fd (vs ≤1pw)	-	-	-	-	-
(1-2pw)	0.002944	0.001904	1.55	0.12	(-0.0008,0.0067)
(3+pw)	0.005576	0.002514	2.22	0.03	(0.0006,0.0105)
FrtVeg (vs 5+pd)	-	-	-	-	-
(≤1pd)	0.007114	0.003249	2.19	0.03	(0.0007,0.0135)
(1-3pd)	0.006505	0.002267	2.87	0.00	(0.0021,0.0109)
(3-5pd)	0.003028	0.002241	1.35	0.18	(-0.0014,0.0074)
RecPA (vs 0 hpw)	-	-	-	-	-
(≤1 hpw)	-0.00441	0.002531	-1.74	0.08	(-0.0094,0.0005)
(1-2 hpw)	-0.00854	0.002508	-3.4	0.00	(-0.0135,-0.0036)
(3+ hpw)	-0.01691	0.002506	-6.75	0.00	(-0.0218,-0.0120)
Smoking (vs current)	-	-	-	-	-
Ex-Regular	0.00805	0.008522	0.94	0.35	(-0.0087,0.0248)
Never	0.002916	0.006702	0.43	0.66	(-0.0102,0.0161)
Age	0.001886	0.000127	14.88	0.00	(0.0016,0.0021)
age*ex-reg smoke	-0.00035	0.000171	-2.03	0.04	(-0.0007,0.0000)
age*nvr smoked	-0.00032	0.00015	-2.14	0.03	(-0.0006,0.0000)
BMI	0.008833	0.000194	45.43	0.00	(0.0085,0.0092)
EducL (vs higher)	-	-	-	-	-
(below higher)	0.00647	0.001952	3.31	0.00	(0.0026,0.0103)
(No quals)	0.008006	0.002606	3.07	0.00	(0.0029,0.0131)
(student)	-0.00337	0.004252	-0.79	0.43	(-0.0117,0.0050)
Constant term	0.602803	0.008012	75.23	0.00	(0.5871,0.6185)

Table 7.3: Coefficients obtained from Generalized Linear Modelling of factors influencing waist to hip ratio (Males)

	Coef.	Std. Err.	*z*	P>*z*	95% Conf. Interval
Snacks (vs <1pw)	-	-	-	-	
(1-2pw)	-0.00623	0.002881	-2.16	0.031	(-0.0119,-0.0006)
(3-5pw)	-0.00514	0.002753	-1.87	0.062	(-0.0105,0.0003)
(5+pw)	-0.00979	0.003195	-3.06	0.002	(-0.0161,-0.0035)
FrtVeg (vs 5+pd)	-	-	-	-	
(<1pd)	0.011467	0.004014	2.86	0.004	(0.0036,0.0193)
(1-3pd)	0.00391	0.00256	1.53	0.127	(-0.0011,0.0089)
(3-5pd)	0.004394	0.002434	1.8	0.071	(-0.0004,0.0092)
RecPA (vs 0 hpw)	-	-	-	-	
(<1 hpw)	0.00129	0.002838	0.45	0.65	(-0.0043,0.0069)
(1-2 hpw)	-0.00304	0.002665	-1.14	0.254	(-0.0083,0.0022)
(3+ hpw)	-0.00967	0.003165	-3.06	0.002	(-0.0159,-0.0035)
Smoking (vs current)	-	-	-	-	
Ex-Regular	-0.01348	0.002901	-4.65	0	(-0.0192,-0.0078)
Never	-0.01977	0.002517	-7.86	0	(-0.0247,-0.0148)
Age	0.001269	0.000081	15.66	0	(0.0011,0.0014)
BMI	0.004749	0.000177	26.81	0	(0.0044,0.0051)
EducL (vs higher)	-	-	-	-	
(below higher)	0.003683	0.002359	1.56	0.119	(-0.0009,0.0083)
(No quals)	0.01181	0.003219	3.67	0	(0.0055,0.0181)
(student)	-0.00323	0.006214	-0.52	0.603	(-0.0154,0.0089)
NSSEC (vs Prof)	-	-	-	-	
(Intermediate)	0.001432	0.002606	0.55	0.583	(-0.0037,0.0065)
(Routine)	0.000344	0.002433	0.14	0.887	(-0.0044,0.0051)
(LT unemployed)	0.025099	0.009147	2.74	0.006	(0.0072,0.0430)
(Other)	-0.00795	0.019841	-0.4	0.689	(-0.0468,0.0309)
Constant term	0.64139	0.006927	92.59	0	(0.6278,0.6550)

Table 7.4: Coefficients obtained from Generalized Linear Modelling of factors influencing waist to hip ratio (Females)

nificant predictors. As in males reduced fruit and vegetable intake is associated with an increased WHR. Low snack intake results in a significantly higher WHR. Never smokers and ex-regular smokers also tend to have a lower WHR than current smokers, however no interaction is observed with age. Long term unemployed and those with no qualifications have a significantly higher WHR than the managerial/professional class and those with higher education respectively.

7.5 Discussion

This study examined the determinants of WHR in 2006 HSE data using two contrasting techniques. Edge relation features derived from Bayesian Model Averaging were compared with results from the standard epidemiological method of generalized linear modelling (GLM).

Both techniques identified BMI and Age as the major determinants of WHR- in agreement with existing literature [192, 193]. This was true of both males and females. Cross tabulations of WHR given age and BMI can be seen in tables 7.5 and 7.6. Overall GLM was shown to be a significantly more sensitive technique than use of Bayesian network edge relation features at identifying factors relevant to WHR. This is unsurprising given that the Bayesian approach was not specifically examining WHR determinants, but fitting a model to the joint distribution of all variables. Secondly, the inability of the BN topology model to incorporate continuous data results in the loss of large amounts of information, making linear relationships harder to detect. When two nodes are a parent of another, their output levels are multiplied to produce a number of new categories, each of which is independent, meaning the effects of each node are not resolved independently. In addition, the assigned pseudocount may penalise nodes from having numerous parents. Although a pseudocount of 1.0 is the minimum observable prior, where two nodes with several outcome levels are parents the number of input levels of the child may be large, and a prior of 1.0 may make a significant contribution to the evidence. Selection of a smaller prior- (e.g. 0.01) may result in a less peaked and more easily traversable distribution. The combination of discrete and continuous nodes in the same network may provide a solution to these problems; however this results in the likelihood function being non-conjugate with the prior, meaning the posterior probability is not easily expressed in closed form. Consequently, exact calculation would not be possible, and approximate techniques would require prohibitively long calculation times.

The results from the GLM for all categories indicate that there are numerous predictors of WHR other than age and BMI- most notably smoking, which results in an increased WHR of 0.01977 for current smokers compared to never smokers in females (table 7.4) and an additional 0.0029 per year of life in males (table 7.3). Where a variable is a weaker predictor of another, we may expect to observe an edge between the two nodes infrequently. However, this is not the case in the observed data- only BMI and age edges are ever observed with WHR. This suggests that the edge relation feature is not necessarily a reliable indicator of association or predictive value. This may be a result of the highly peaked probability distribution only allowing transition between a relatively small group of DAGs, which do not include the specified edge. As the dataset size increases the probability landscape increasing resembles a Dirac delta distribution, with a small number of topologies becoming dominant. Consequently, this approach does not appear sensitive to subtle effects.

Irrespective of this issue, Bayesian Model Averaging represents a useful modelling tool. The chief advantages of this approach are that it is data-driven and provides an intuitive visualisation of the structure within the dataset. A data-driven approach removes investigator bias and model selection bias associated with hypothesis led research. However, it is recognised that the choice of variables included in the model is not unbiased- further applications of the technique may use more variables in a less structured manner.

Correlated data can cause problems within GLM analyses, especially where the purpose is to identify dependent relationships. Correlated covariates can lead to inflated standard error estimates [208, 209] and can be misleading during the model selection process. For example, in the GLM model, in males cake intake is identified as a predictor of WHR following the stepwise regression procedure. In the female data snack intake is identified as a predictor but cake intake is not. In both cases these intake indicators are interchangeable, both variables generate a significant P-value (data not shown). In a situation where the investigator was not aware of any correlation between them, they may erroneously conclude that one was a significant predictor and the other wasn't. Awareness of potential confounders within the dataset is important when performing GLM- this approach provides a visualisation of the dataset that can help highlight potential confounding factors. Both techniques are of course susceptible to confounders not observed within the dataset, the presence of such factors is suggested by the identification of social class and education as predictors of WHR. Highly correlated covariates

make it extremely difficult to identify potential causal relationships, BMA adds new utility by showing dependencies between *all* variables.

In the exploratory stages of this analysis, data was stratified by age categories to examine differences in determinants between age groups. Metropolis Hastings sampling of Bayesian network topologies was performed on each strata. However data strata are subject to correlation collision [75], *i.e.* the phenomenon of another variable acting as a proxy for the variable on which the data was stratified. This leads to misleading results. In this study several variables exhibited close dependency relationships with WHR due to this bias. As the model highlighted the close correlations between age and several socio-demographic variables from the main model, this was easily identified. Visualisation of data structure is a useful tool in avoiding such pitfalls.

From this analysis it is clear that BMA in its current form is not a viable alternative to GLM when investigating determinants of a single variable; the lack of sensitivity associated with categorical data and prior selection make it impractical. However BMA has significant strengths; particularly with complex datasets where multiple conditional dependencies are present. BMA therefore adds most value as a data-visualisation tool for shaping hypotheses. It may therefore be a useful adjunct to conventional regression modelling.

The relevance of epidemiological findings from this study are limited due to the lack of longitudinal data. Cross sectional studies fail to capture the relevant lifetime exposures and lifetime body fat deposition. The data in this study are not representative of the general population- the HSE are prone to participation bias, restricting the generalisation of results to the population as a whole. The data relies on self reporting of food intake, and agreement to provide weight measurements, both of which are known to be unreliable, and linked with weight status [131, 132]. However, the study provides a useful illustration of the limitations of the BMA technique in this typical epidemiological context.

(a) Males

Age	Waist Hip Ratio						
	≤ 0.80	0.80 – 0.85	0.85 – 0.90	0.90 – 0.95	0.95 – 1.00	≥ 1.00	Total
16-24	85	108	81	43	10	4	331
	25.7	*32.6*	*24.5*	*13.0*	*3.0*	*1.2*	
25-34	29	88	159	117	46	20	459
	6.3	*19.2*	*34.6*	*25.5*	*10.0*	*4.4*	
35-44	11	79	196	234	160	68	748
	1.5	*10.6*	*26.2*	*31.3*	*21.4*	*9.1*	
45-54	5	33	136	214	177	104	669
	0.7	*4.9*	*20.3*	*32.0*	*26.5*	*15.5*	
55-64	7	19	106	222	214	192	760
	0.9	*2.5*	*13.9*	*29.2*	*28.2*	*25.3*	
60-74	1	9	29	77	102	96	314
	0.3	*2.9*	*9.2*	*24.5*	*32.5*	*30.6*	
Total	138	336	707	907	709	484	3,281
	4.2	*10.2*	*21.5*	*27.6*	*21.6*	*14.8*	

(b) Females

Age	Waist Hip Ratio						
	≤ 0.70	0.70 – 0.75	0.75 – 0.80	0.80 – 0.85	0.85 – 0.90	≥ 0.90	Total
16-24	31	71	66	40	12	10	230
	13.5	*30.9*	*28.7*	*17.4*	*5.2*	*4.3*	
25-34	30	131	179	126	67	25	558
	5.4	*23.5*	*32.1*	*22.6*	*12.0*	*4.5*	
35-44	30	144	278	248	140	68	908
	3.3	*15.9*	*30.6*	*27.3*	*15.4*	*7.5*	
45-54	17	81	181	237	179	92	787
	2.2	*10.3*	*23.0*	*30.1*	*22.7*	*11.7*	
55-64	8	71	173	258	196	152	858
	0.9	*8.3*	*20.1*	*30.1*	*22.8*	*17.7*	
60-74	3	17	47	92	84	78	321
	0.9	*5.3*	*14.6*	*28.7*	*26.2*	*24.3*	
Total	119	515	924	1001	678	425	3,662
	3.2	*14.1*	*25.2*	*27.3*	*18.5*	*11.6*	

Table 7.5: Cross tabulation of WHR by Age groups, percentages in *italics*

(a) Males

BMI	Waist Hip Ratio						
	≤ 0.80	0.80 – 0.85	0.85 – 0.90	0.90 – 0.95	0.95 – 1.00	≥ 1.00	Total
≤ 19.0	34	35	18	9	0	1	97
	35.1	*36.1*	*18.6*	*9.3*	*0.0*	*1.0*	
19-24.9	88	210	316	198	63	10	885
	9.9	*23.7*	*35.7*	*22.4*	*7.1*	*1.1*	
25-29.9	16	86	330	523	352	129	1,436
	1.1	*6.0*	*23.0*	*36.4*	*24.5*	*9.0*	
30-34.9	0	5	37	159	245	225	671
	0.0	*0.7*	*5.5*	*23.7*	*36.5*	*33.5*	
35-39.9	0	0	3	17	41	83	144
	0.0	*0.0*	*2.1*	*11.8*	*28.5*	*57.6*	
≥ 40	0	0	3	1	8	36	48
	0.0	*0.0*	*6.3*	*2.1*	*16.7*	*75.0*	
Total	138	336	707	907	709	484	3,281
	4.2	*10.2*	*21.5*	*27.6*	*21.6*	*14.8*	

(b) Females

BMI	Waist Hip Ratio						
	≤ 0.80	0.80 – 0.85	0.85 – 0.90	0.90 – 0.95	0.95 – 1.00	≥ 1.00	Total
≤ 19.0	15	59	63	28	6	1	172
	8.7	*34.3*	*36.6*	*16.3*	*3.5*	*0.6*	
19-24.9	81	308	446	359	133	43	1370
	5.9	*22.5*	*32.6*	*26.2*	*9.7*	*3.1*	
25-29.9	16	123	288	363	269	130	1189
	1.3	*10.3*	*24.2*	*30.5*	*22.6*	*10.9*	
30-34.9	5	18	84	160	166	140	573
	0.9	*3.1*	*14.7*	*27.9*	*29.0*	*24.4*	
35-39.9	1	5	29	61	75	78	249
	0.4	*2.0*	*11.6*	*24.5*	*30.1*	*31.3*	
≥ 40	1	2	14	30	29	33	109
	0.9	*1.8*	*12.8*	*27.5*	*26.6*	*30.3*	
Total	119	515	924	1001	678	425	3,662
	3.2	*14.1*	*25.2*	*27.3*	*18.5*	*11.6*	

Table 7.6: Cross tabulation of WHR by BMI groups, percentages in *italics*

Chapter 8

Discussion

8.1 Overview

The introduction to this thesis outlined the need to move towards a more complete method of data modelling, in order to incorporate the complexity present within systems. Obesity is a complex, multifaceted problem that has been linked with a range of socio-demographic factors, although importantly these are correlations, and not necessarily causes. Socio-demographic variables are inherently highly correlated [74], which can be problematic for models that seek to explain variation. Graphical models have the advantage of coping well with highly correlated data, as they model the joint distribution of the system rather than the outcome of a single variable [139]. In addition, graphical models can be intuitively understood as dependence relationships are represented visually by arcs between nodes. Complex patterns of multi variable interaction can be understood far more easily than the equivalent output from a regression model.

Although they have proved useful in related fields, machine learning techniques are rarely applied in epidemiology. This is unsurprising when epidemiologists largely seek independent causes of disease, whereas strengths of machine learning tend to lean more towards classification and pattern recognition applications [210]. The body of work in this thesis represents an attempt to use machine learning techniques to identify conditional dependencies present in typical epidemiological datasets. Crucially, this work is not hypothesis led, but data driven. The analyses carried out are exploratory, intended to identify interesting relationships that may be able to inform the design and targeting of interventions. With increasing availability of data, such as centrally held medical records with linked personal information such as postcodes, the argument to utilise such minimally supervised machine learning techniques is compelling. I hope that the work carried out in this thesis not only provides results directly relevant to the study of obesity, but also that it may inform wider endeavours in emerging field of machine learning in epidemiology. This discussion section is broken into two parts; the first details the results of the analyses and considers the relevance of the work to more general problems in obesity. The second section is of a much more technical nature, and discusses the limitations of the method and explores the potential for future development of the technology.

8.2 Relevance to Obesity

Challenges in Obesity

Obesity is an important public health problem that requires substantial investment from health services [15]. Although often laid at the door of individual responsibility, changes in the environment have resulted in a shift in behavioural habits over the past few decades. Greater availability and affordability of energy dense food has led to higher consumption, and changes in the labour market and convenience technologies have resulted in less occupational activity and more sedentary use of leisure time. Despite the apparently simple nature of the problem, obesity epidemiology presents several challenges. A major obstacle to policy making is the lack of interventions proven to reduce or even restrain obesity at the population level. The identification of effective interventions is limited by a lack of knowledge of the obesogenic environment, a better understanding of which is required to identify relationships that may lead to the design of new interventions, or the localisation of population subgroups where they can be effectively targeted. However, the complexity and the highly correlated nature of associated factors mean that the utility of standard epidemiological techniques is limited.

Thesis Summary

In this thesis Bayesian networks were used to model relationships between variables in datasets from Health Surveys for England (HSE) data. A Bayesian network (BN) is the graphical representation of a probability distribution, where conditional dependencies between variables are represented by arcs between nodes. A BN is composed of two components, the structure of nodes and arcs; and parameters, the probabilities associated with these arcs. Given data, it is possible to score the probability of a Bayesian network structure, *independently* of parameters. The structure of a BN imparts information about conditional dependencies present. A computational technique, Metropolis Hastings (MH) sampling is applied to estimate the posterior probability of the existence of arcs between nodes given a dataset.

The first results chapter (Chapter 5) examined the relationships between a variety of obesity related factors. The dataset included a range of socio-demographic variables and several indicators of both energy intake and expenditure. Data from the 2003 and 2006 HSE were examined, with data stratified by gender. MH sampling was performed on the four datasets and a sample of the distribution of net-

work topologies was derived for each. Results were plotted in a graphical format enabling easy comprehension of complex data patterns. Broadly, conditional dependencies were strong between socio demographic variables, and between indicators of energy intake. Age and ethnicity exhibited strong conditional dependencies with dietary intake, and fruit and vegetable intake showed a firm relationship with social class. These effects were observed in both males and females. Network structures indicated different determinants of recreational physical activity (RPA) between males and females. In males, health and age were associated with RPA, while in females age was replaced by education level. Although age and education level are closely correlated, as indicated by the consistently observed arc between them, this difference in topology suggests a different dynamic as detailed in the section on d-separation (3.1.2). The topology observed indicates that RPA in females is independent of age given education level, and in males RPA is independent of education level given age. Although in both cases there is likely to be some residual effect of the other variable (possible reasons for this are discussed in detail later), it remains that the primary determinant is different. An analysis of determinants of RPA was carried out on the 2003 HSE by Stamatakis et al [170]; although it was noted that age and various socio-demographic factors were implicated in determining RPA levels, this potentially informative relationship was not reported. Results from this analysis held relatively consistent between years, suggesting real rather than artefactual effects. The results derived from this study may be of use to policy makers; an approach to increase RPA in males may concentrate on encouraging participation in older groups, and arresting the observed decline in age (see fig. 5.17(a)). Interventions in females may target social barriers to participation such as cost and access. The relationship between social class and fruit and vegetable intake is well documented [59, 169] and reinforces the need to improve dietary quality in these groups. The results do not imply any causal relationship however. Although individuals in lower social classes consume fewer fruits and vegetables, this highlights the target of intervention rather than a solution. Policymakers may also note the relationship between ethnicity and dietary intake. Further research may identify the reasons for the differences reported.

Following the exploratory approach of Chapter 5, a more practical problem was addressed in the second results chapter (6). Survey data is prone to bias, particularly participation bias, meaning published rates of health behaviour derived from such studies are likely to be inaccurate. Participating individuals are likely to be healthier, whiter, and wealthier [129] than the population which they are in-

tended to represent. Using a Bayesian framework I built a model of obesity related behaviours given socio-demographic factors from HSE data. The model was then combined with matched socio-demographic data from a more representative population dataset. Essentially, the model is a Bayesian classifier to which data from the second dataset is applied. A Bayesian approach is well suited due to the ability to simultaneously model numerous outputs, and to account for the uncertainty generated by small groups. If a group is rare in the HSE sample, but common in the census data, small numbers will fail to overwhelm the priors and will not skew the data. The intended use of this application is to enable policymakers and public health professionals to estimate the number of individuals in a population that meet a condition (or conditions) of interest. For example, the model estimated that 61.9% of the $904,754$ females in the Greater Manchester population participated in no recreational physical activity. Furthermore, it is possible to calculate this within groups, so that the health behaviour of any subgroup (e.g. by social class or ethnicity) can be estimated. Currently the model is implemented in R code, and is not appropriate for non-expert use in its current form. However, with further development it would be possible to introduce a graphical user interface (GUI), which would allow a user to select the dataset, subgroups and behaviours of interest. Efforts to perform similar tasks have previously been reported. Molitor *et al* [187] built a model of birth weight given area water quality and applied data collected at a much higher level to estimate effects of trihalomethanes. This approach used Bayesian networks to combine a set of non-Bayesian submodels. Despite using Bayesian networks the methodology is distinct to that described here, however this does highlight the potential usefulness of BNs in this application.

A single outcome measure for obesity fails to address the full variation of health risks associated with adiposity. The most common indicator, the Body Mass Index (BMI) is useful but limited. The distribution of body fat has a large bearing on health risks. Waist to Hip Ratio (WHR) is a proxy indicator of the proportion of visceral fat, adipose tissue that surrounds the vital organs. This is predictive of cardiovascular disease independently of BMI [11]. However, determinants of body fat distribution remain elusive. The third results chapter aims to exploit the direct dependence properties of Bayesian networks to help uncover determinants of body fat distribution. Similar methods were applied to that used in chapter 5; a Metropolis Hastings sampler was implemented over the space of Bayesian network topologies representing a dataset including physical measurements. Mixing was better than that observed in Chapter 5, possibly as a result of the lower number of

individuals included in the study. A generalised linear model (GLM) was applied to the same data to highlight differences between the approaches. The Bayesian topology approach successfully identified the two major determinants of WHR, age and BMI, however it lacked the sensitivity of the GLM, possible reasons for the lack of sensitivity are discussed later in this chapter. Despite this, the approach still has significant utility in terms of its ability to identify relationships between *all* variables with visual clarity.

The work conducted throughout this thesis has wider implications than the results of the individual analyses. Machine learning techniques are likely to become an important tool for epidemiologists. As demonstrated by the results of Chapter 5, inference of relationships from Bayesian networks structure can yield interesting and potentially useful results. This analysis has identified a relationship that has not been identified by conventional methods [170], providing an excellent example of the utility of this approach.

Causality

Although it is hugely tempting to assign causality in these instances, it must be remembered that we are dealing with a set of correlated variables derived from a snapshot of a highly complex system. The challenge is to separate the dependencies from conditional dependencies. It is likely that socio demographic variables act as proxy indicators for dozens of factors influencing the minor food and exercise choices that individuals make every day. It is unlikely that the social class of an individual will cause them to choose to buy processed food over fresh produce, however there are a number of related factors such as disposable income, accessibility of markets, and food awareness/attitude. Identification of the most important generic variables informs which potential *causes* we should study and target. Dozens of these choices are made by individuals every day, whether to take the car, whether to take the stairs or lift, whether to eat fast food. The causality of these many individual decisions cannot be evaluated at such a broad level. The utility of this approach lies in the value of an exploratory journey through a complex data space; the output of which is a series of assertions of conditional dependence that generates or informs a hypothesis, providing potentially fruitful avenues for future research. In epidemiology, we may be preoccupied with identifying which of a set of broad factors explains most of the variation in the response variable (typically the presence or some indicator of disease). This has proved to be successful at identifying relevant factors where a causal mechanism is present. However, when there

is no single mechanism as in obesity, conventional epidemiological modelling is not suited to these analyses. The more holistic approach displayed here allows and encourages an investigator to consider all interdependencies present in a system. This work is not intended as an alternative to standard techniques of regression, but as a data viasualisation and hypothesis shaping tool.

Further Directions of Research

Much of the early work on this the thesis revolved around a literature review of documented attempts to reduce obesity prevalence, with a view to creating a policy model. It was quickly apparent that there was insufficient evidence to inform such a model. Further, there was a lack of realistically complex obesity epidemiology, and Bayesian networks were a potential tool for investigating such complexity. Perhaps the key concept is that Bayesian networks are not causal models, but a method of explaining data. Although BNs can be very good at predicting an individuals behaviour given socio demographic input, this does not mean that it is meaningful to manually adjust the parameters of a Bayesian network to predict impact on health behaviour. The available data is a snapshot of a system, meaning we cannot build a picture of the dynamic relationships between variables, only how they can predict the states of one another.

If we were to create such a policy model of the effects on population behaviour, a model of personal choice is required. The smoking literature provides a number of such models of personal choice which may be exploited [211]. Artificial neural networks or even Bayesian networks are both well suited to modelling individual choice. Unlike smoking, the number of choices made relevant to obesity is enormous, with dozens of choices made every day. This may be problematic, although availability of data is likely to be the largest issue. Currently data capable of informing such models is not available. The closest datasets are likely to be held by private companies such as supermarkets. Consequently, this is not currently a realistic approach. Nonetheless, the research conducted here has begun to unpick some of the complexities in the relationship between obesity and socio-demographic factors.

This thesis has advanced the idea of data driven models within epidemiology, and has tackled many associated technical barriers. I have utilised Bayesian networks to combine representative and non-representative datasets and applied this to predict wider behaviour. Chapters 5 and 7 have provided some potentially useful avenues for policy makers to explore.

8.3 Further Development of Methods

Heuristic Issues

Inferring relationships in data from Bayesian network topologies is a computationally demanding process, that must accede to various approximations and restrictions in order to be tractable. These have been discussed in depth in section 4.6, the heuristic improvements and approximations detailed in these sections may provide a useful resource for investigators wishing to employ similar methods. However, several technical problems were encountered that limited the usefulness of the approach. With greater technical knowledge it may be possible to overcome some of these obstacles. This section discusses such issues and explores the potential for further development of the method.

Throughout the thesis, the use of Bayesian model averaging was frustrated by a lack of sensitivity. Several examples were observed of variables with high connectivity, but with few parents. In the example of recreational physical activity (RPA) in Chapter 5, the structure implied that in females RPA was independent of age given educational status. Closer examination of the data seemed to indicate that although education level appeared to be the dominant predictor, the implication of conditional independence was unreasonable. This failure to acknowledge weaker effects was also observed in the WHR work in chapter 7; a linear regression model identified smoking and alcohol intake as significant predictors of WHR, although no such relationship was identified in the structural learning approach. There are several possible causes for the apparent penalisation of nodes with several parents. The most obvious is that when multiple discrete variables are parents of a node, the number of input levels is the product of the output levels of each parent, each input level is resolved *independently*. There is no concept of ordinality in BN models of discrete data, as a result influence is diluted as more parents are added. This is in contrast to a continous Bayesian network, which exhibits a conditional distribution, allowing one parent to have a much more significant effect than another. The inclusion of another parent does not directly interfere with the influence of existing parents. Furthermore, as the number of input levels becomes greater, so does the influence of the prior in relation to the counts. This is because each input level is associated with a different Dirichlet distribution, and hence additional pseudocounts. In continuous networks inclusion of multiple parents does not result in an exponential rise in the contribution of pseudocounts. As a result it may be expected that

nodes with few outcomes are more likely to be observed as co-parents than nodes with many, this may be a potential source of bias. However, observational data is rarely in entirely continuous form. A solution would be to use a mixed network that allows the interaction of continuous and discrete variables. A mixed network can be constructed, however, it is not straightforward to evaluate the evidence for the network structure given data. The posterior distribution will be highly complex and expression in closed form will not be possible in all cases. Approximation methods are available [212, 213], as are packages that can perform it such as INFER.NET. As the technique of MH sampling relies on approximating a highly complicated space, a further approximation step is too laborious to be practical. Currently the calculation of the network's log evidence (see 4.6.1) takes a fraction of a second, any significant extension to this would render MCMC intractable.

In the analyses performed in this thesis, networks were limited to relatively few nodes when compared with other applications of similar technologies [107, 124], restricting the complexity of networks that could be examined. This limitation was mainly due to concerns associated with successful mixing and convergence over network structures, and the computational load associated with the REV move. When networks become large, implementation of the MCMC technique becomes more difficult. The probability landscape becomes more complex, the number of possible networks increases exponentially with additional nodes, which makes the space of network topologies harder to transverse and hence approximate. As shown in section 4.4.2, without the implementation of the Grzegorczyk-Husmeier REV move, mixing over network topologies is infeasible. As the number of nodes in a network increases, the computational cost of the REV move rises exponentially. The restrictions (section 4.6) that must be introduced to restrain the computational load become more stringent, which increasingly compromises the integrity of the sampler. The REV move was designed for continuous rather than discrete data, which may explain some of the issues experienced. For networks with more than 20 nodes, other approaches may have to be considered.

Ideally a very large number of variables could be included, making the approach more data driven, without the variable selection step that raises concerns of investigator bias. This would reduce the amount of time necessary to choose and prepare the dataset. However, such an approach would raise several problems. The study relies on identifying dependence relationships between variables using BN structure. If the technique were applied to a whole dataset tight clusters of very similar variables would be observed. Much of the MCMC would be spent

maximising the evidence of these features rather than identifying relationships of interest. The dataset would still require cleaning, and the presence of missing data may be problematic. This may lead to unexpected or spurious results such as socio-demographic factors displaying arcs to nodes with high levels of missing data.

In order to reduce the number of discrete variables present in the network, dimension reduction techniques such as Principal Component Analysis (PCA) or Factor Analysis may be implemented [214]. These are relatively common approaches in machine learning [210]. However, in cases where we wish to infer something meaningful from the relationships present in the data as opposed to the solution of a classification problem, replacement with 'pseudo-variables' may not be appropriate. Nonetheless, latent classes may exist in the data. A latent variable is a variable that is unobserved, which may correspond to a real variable that is not easily measured or a more abstract concept. In this data latent class analysis may reveal groups such as 'health conscious' which determines behaviour, and is informed by socio-demographic variables. This would be relatively straightforward to implement by including a node with all values missing. However incorporation of missing variables into a Bayesian network requires approximation techniques that are intractable in MCMC analysis.

MCMC Mixing over Bayesian Network Topologies

Mixing is probably the most important factor that determines the practical network size that can be evaluated. Successful mixing and convergence was not achieved in all instances in this thesis. The failure of several chains to converge on the most likely regions indicates that this work pushes the limits of what datasets Metropolis Hastings (MH) sampling can be reasonably applied to. To an extent, this is an unavoidable consequence of exploring a complex discrete space, however the high number of datapoints exacerbates this issue. Application of this technique to larger datasets is probably not feasible due to mixing concerns. Numerous heuristic variations of MH sampling techniques exist to improve mixing, some extensions are available that were not pursued in this thesis. These are discussed briefly in section 4.2, and more thoroughly in several resources (e.g. [215]).

Incorporating Investigator Knowledge

Throughout this thesis, data has been stratified to investigate different effects of population subgroups. In Chapter 5, the different dynamics of obesity in males and

females were examined. Stratification of data allows examination of patterns in different subgroups which may be informative. When a strong effect is present in a small group, an arc may not be observed as relatively few individuals are in this subgroup. For example in young individuals access to transport may be a highly powerful indicator of incidental physical activity, but less so in older groups. Although there will be a highly significant effect at one input level, as the majority of input levels are uninformative, an arc is unlikely. This is not too surprising, as any method will struggle to detect masked effects. The Bayesian topology method identifies the arcs that influence the evidence of the whole dataset, which is a relatively crude indicator of influence. A distinct subgroup may have very different determinants of health behaviour. This may be an argument for the inclusion of latent classes (discussed above). Linear regression has the capacity to specifically fit an interaction term; which is not possible in this context. The extension of a Bayesian network such that input levels can be merged, allowing the existence of 'sub-arcs' may be worth pursuing. During the work conducted during this thesis, splitting data into separate subgroups was attempted. Initially this had the effect of attaching very high importance to nodes that had high correlations with the stratified variable, presumably as the result of collision bias [75]. The probability landscape of network topologies is much flatter and results in much freer mixing. However, without knowledge of true subgroups, analyses may be prone to spurious results. The technology described is best suited to identifying consistent rather than subtle relationships in complex data.

In some applications of Bayesian networks it may be desirable to incorporate expert knowledge. For example an expert may have evidence of a direct causal relationship that should be included in the model. Prior belief is directly incorporated into Bayesian systems. In the Bayesian networks described in this thesis, two classes of prior are used. Firstly, priors on counts (*i.e.* pseudocounts), the α in the marginal likelihood of the data shown in eq. 8.1; the second is the prior on network structure, for which I use the complexity penalising prior in eq. 8.2.

$$\Pr(D|H) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}. \tag{8.1}$$

$$\Pr(H) = \frac{1}{\Pi} \prod_{n=1}^{N} \binom{N-1}{|\pi_n|}^{-1}. \tag{8.2}$$

Pseudocounts act as they are named, as unobserved counts. If an investigator

wishes to incorporate an opinion that a particular state of node *A* has an influence on node *B*, then this must be represented by assigning pseudocounts. The difficulty is associated with taking a qualitative expression of certainty from an individual and converting it into a figure. The pseudocount can only exist where *A* is a parent of *B*. Assignment of pseudocounts **will** have an effect on structure, but the precise nature of this effect is difficult to predict. In order to assign prior knowledge of structure, the prior in 8.2 is manipulated. The simplest method of doing this is to add a penalty on networks that meet or fail to meet a certain condition, such as an arc between *A* and *B*. Again the complexity derives from converting an often verbal expression of certainty into numeric form.

In Chapter 6, a specialised scoring criterion was used to identify determinants of the set of indicators of energy intake and expenditure. The usual method of calculating a log likelihood is the product of all marginal likelihoods of each node input combination (eq. 8.1) and the prior on the topology (eq. 8.2). The specialised criterion limited this to the set of indicator nodes, due to the lack of interest in relationships between socio-demographic variables. Although this was a useful approach for the task at hand, it is not informative for us to investigate influences on one node. This would essentially be a naïve Bayes classifier. Regression approaches are more appropriate in this case.

Assigning Significance

Well established techniques such as regression models have an easily understandable and reportable figure to explain the strength and confidence in an effect. This is the significance measure, or casually 'The probability of observing the relationship by chance'- the probability of a type I error. The natural alternative to the significance measure is the edge relation feature (ERF). This represents an estimate of the probability that the arc is present over the posterior distribution of network topologies. If the Metropolis Hastings algorithm provides representative and independent samples from the distribution then a confidence interval for the proportion can be generated. The Wilson score interval approximation [216] is probably most appropriate here, due to the high number of edges that tend to bimodal probabilities of zero and one. In theory this proportion estimate could be tested against any proportion using some test of significance. However, the utility of such a measure may be questionable. The ERF doesn't necessarily indicate confidence in the conclusions, it provides a a strong indication of the the confidence in the structure of the best networks. More interesting perhaps is the sensitivity of the results to perturbations

in the data. This can be evaluated by bootstrapping. Bootstrapping is a resampling method for evaluating an estimator (such as a confidence interval) that is hard to evaluate analytically. It is not tractable to perform bootstrapping on results derived from a Metropolis Hastings algorithm, as the MH algorithm would have to be performed on each data sample. It is however possible to bootstrap the optimisation algorithm, this facility was implemented in the C# program.

In closing, I would like to briefly revisit the stated aim of the thesis:

> 'To apply Bayesian networks to typical problems in obesity epidemiology and to evaluate their utility in this context.'

This work represents an attempt to answer some useful and pressing questions within obesity research using a technology new to the field. The results described here have shown the capability and limitations of Bayesian network methods in the context of epidemiology. Specifically, I have identified relationships between obesity related factors that may be informative to policy, produced estimates of health behaviour in real populations, and lastly highlighted potential determinants of body fat deposition. This thesis has demonstrated the advantages of graphical models, both methodologically and visually, and applied machine learning techniques in epidemiology. Many methodological barriers have been encountered and overcome; it is hoped that these heuristics may be of use to any future researchers. The methodology employed in Chapter 6 is an innovative method that uses Bayesian networks to combine datasets and has numerous applications.

As epidemiology moves towards the study of more complex disease phenotypes such as diabetes, obesity and cancer, individual risk factors become less apparent with problems of correlation. Complex diseases may interact with one another, which further complicates identification of causal factors. With identification of single risk factors becoming less relevant to the study of disease, graphical modelling techniques that are capable of modelling complete systems are likely to be of increasing importance. However, the results described here are conducted on single time point data where a state of one variable may infer the likely state of another. Extension of the methodology is required before these modelling techniques can be applied to longitudinal data. This is likely to be non-trivial. Numerous other limitations persist, such as the ability to combine discrete and continuous data, and mixing concerns with larger datasets. Further methodological advancement may

be needed to perform similar analyses on a larger scale.

Through this thesis, I have shown the current utility of Bayesian networks in epidemiology. However this utility is limited by the complexity and awkwardness of the method, as well as sensitivity issues and loss of information in categorical data. Irrespective of the precise nature of future research involving Bayesian networks in epidemiology, it is hoped that this thesis may provide information of value to current and future investigators.

8.4 Evaluation

This thesis is the end product of a journey that has taken over three and a half years. There have been several dead ends, and frustrations, but equally many valuable lessons learned. The initial aim of this thesis was to build a system that modelled obesity in populations, and provided a tool allowing health professionals to model potential effects of interventions. This proved too difficult, mainly due to the size and vagueness of this proposition. The main challenge I faced was to turn a series of loose ideas into a set of concrete research questions.

Much of the early work on the thesis revolved around a thorough literature review of documented attempts to reduce obesity prevalence. This work proved to be of limited value when the more technical nature of the PhD emerged, my time may have been better spent familiarising myself with Bayesian statistics or programming methods. However, this work did prove valuable in helping me to see the larger context and complexity of the obesity problem, which I certainly believe added substantially to the work.

Looking back on the achievements of the PhD I am pleased to have developed and applied a technique that has potentially useful applications, and to have used it to obtain interesting results. I am convinced it provides value in hypothesis generation and data investigation in epidemiological datasets. However, I am less convinced that it has the necessary characteristics to become widely adopted within the field. The implementation of the Bayesian averaging technique is extremely labourious and requires substantial computer processing time. In addition, the interpretation of the results requires a trained eye and a clear understanding of the underlying processes. In the (relatively common) event of an MCMC chain failing to mix, it is vital that this is spotted and rectified by the investigator, otherwise results will be meaningless. Also the difficulty in assigning significance limits its potential to epidemiologists. Despite this, this body of work represents a thorough

exploration of the applicability of an under-used technique to epidemiology.

Bibliography

[1] WHO. Obesity: preventing and managing the global epidemic. Report of a WHO consultation, 2000.

[2] *Statistics on Obesity, Physical Activity and Diet: England*. NHS: The Information Centre, 2009.

[3] J. O. Hill and J. C. Peters. Environmental contributions to the obesity epidemic. *Science*, 280(5368):1371–4, 1998.

[4] Food and drink federation hot issues: Obesity. `http://www.fdf.org.uk/hot_issue_obesity.aspx`, October 2007.

[5] S. A. French, M. Story, and R. W. Jeffery. Environmental influences on eating and physical activity. *Annu Rev Public Health*, 22:309–35, 2001.

[6] World health organisation: European charter on counteracting obesity, 2007.

[7] C. B. Ebbeling, D. B. Pawlak, and D. S. Ludwig. Childhood obesity: public-health crisis, common sense cure. *Lancet*, 360(9331):473–82, 2002.

[8] D. Canoy and I. Buchan. Challenges in obesity epidemiology. *Obes Rev*, 8 Suppl 1:1–11, 2007.

[9] A. Peeters, J. J. Barendregt, F. Willekens, J. P. Mackenbach, A. Al Mamun, and L. Bonneux. Obesity in adulthood and its consequences for life expectancy: a life-table analysis. *Ann Intern Med*, 138(1):24–32, 2003.

[10] Eugenia E Calle, Michael J Thun, Jennifer M Petrelli, Carmen Rodriguez, and Clark W Heath. Body-Mass Index and Mortality in a Prospective Cohort of U.S. Adults. *New England Journal of Medicine*, 341(15):1097–1105, 1999.

[11] D. Canoy, S. M. Boekholdt, N. Wareham, R. Luben, A. Welch, S. Bingham, I. Buchan, N. Day, and K. T. Khaw. Body fat distribution and risk of coronary heart disease in men and women in the european prospective investigation into cancer and nutrition in norfolk cohort: a population-based prospective study. *Circulation*, 116(25):2933–43, 2007.

[12] A. G. Renehan, I. Soerjomataram, M. Tyson, M. Egger, M. Zwahlen, J. W. Coebergh, and I. Buchan. Incident cancer burden attributable to excess body mass index in 30 european countries. *Int J Cancer*, 126(3):692–702, 2010. Renehan, Andrew G Soerjomataram, Isabelle Tyson, Margaret Egger, Matthias Zwahlen, Marcel Coebergh, Jan Willem Buchan, Iain Research Support, Non-U.S. Gov't United States International journal of cancer. Journal international du cancer Int J Cancer. 2010 Feb 1;126(3):692-702.

[13] A. Anandacoomarasamy, I. Caterson, P. Sambrook, M. Fransen, and L. March. The impact of obesity on the musculoskeletal system. *Int J Obes (Lond)*, 2007.

[14] D. W. Haslam and W. P. James. Obesity. *Lancet*, 366(9492):1197–209, 2005.

[15] House of Commons Select Committee. Obesity: Third Report of Session 2003-04. `http://www.publications.parliament.uk/pa/cm200304/cmselect/cmhealth/23/23.pdf`, 2004.

[16] P. R. Nader, M. O'Brien, R. Houts, R. Bradley, J. Belsky, R. Crosnoe, S. Friedman, Z. Mei, and E. J. Susman. Identifying risk for obesity in early childhood. *Pediatrics*, 118(3):e594–601, 2006.

[17] R. C. Whitaker, J. A. Wright, M. S. Pepe, K. D. Seidel, and W. H. Dietz. Predicting obesity in young adulthood from childhood and parental obesity. *N Engl J Med*, 337(13):869–73, 1997.

[18] J. M. Chan, E. B. Rimm, G. A. Colditz, M. J. Stampfer, and W. C. Willett. Obesity, fat distribution, and weight gain as risk factors for clinical diabetes in men. *Diabetes Care*, 17(9):961–9, 1994.

[19] G. A. Colditz, W. C. Willett, A. Rotnitzky, and J. E. Manson. Weight gain as a risk factor for clinical diabetes mellitus in women. *Ann Intern Med*, 122(7):481–6, 1995.

[20] J. B. Meigs, P. W. Wilson, C. S. Fox, R. S. Vasan, D. M. Nathan, L. M. Sullivan, and R. B. D'Agostino. Body mass index, metabolic syndrome, and risk of type 2 diabetes or cardiovascular disease. *J. Clin. Endocrinol. Metab.*, 91:2906–2912, Aug 2006.

[21] C. L. Hart, D. J. Hole, D. A. Lawlor, and G. Davey Smith. How many cases of Type 2 diabetes mellitus are due to being overweight in middle age? Evidence from the Midspan prospective cohort studies using mention of diabetes mellitus on hospital discharge or death records. *Diabet. Med.*, 24:73–80, Jan 2007.

[22] K. M. Narayan, J. P. Boyle, T. J. Thompson, S. W. Sorensen, and D. F. Williamson. Lifetime risk for diabetes mellitus in the united states. *Jama*, 290(14):1884–90, 2003.

[23] O. Pinhas-Hamiel and P. Zeitler. Acute and chronic complications of type 2 diabetes mellitus in children and adolescents. *Lancet*, 369(9575):1823–31, 2007.

[24] O. Pinhas-Hamiel and P. Zeitler. Clinical presentation and treatment of type 2 diabetes in children. *Pediatr Diabetes*, 8 Suppl 9:16–27, 2007.

[25] a Bagust, P K Hopkinson, W Maier, and C J Currie. An economic model of the long-term health care burden of Type II diabetes. *Diabetologia*, 44(12):2140–55, December 2001.

[26] P. Hogan, T. Dall, and P. Nikolov. Economic costs of diabetes in the us in 2002. *Diabetes Care*, 26(3):917–32, 2003. American Diabetes Association Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. United States.

[27] National Statistics: Mid Year Population Estimates. `http://www.statistics.gov.uk/statbase/product.asp?vlnk=15106`, 2009.

[28] S. Wild, G. Roglic, A. Green, R. Sicree, and H. King. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care*, 27:1047–1053, May 2004.

[29] W. A. Davis, M. W. Knuiman, D. Hendrie, and T. M. Davis. The obesity-driven rising costs of type 2 diabetes in australia: projections from the fremantle diabetes study. *Intern Med J*, 36(3):155–61, 2006.

[30] Statistics Great Britain. Office for National. Mortality statistics. deaths registered in. *Mortality statistics. Deaths registered in ... .*, 2006.

[31] B Unal, J A Critchley, and S Capewell. {M}odelling the decline in coronary heart disease deaths in {E}ngland and {W}ales, 1981-2000: comparing contributions from primary prevention and secondary prevention. *BMJ*, 331:614, September 2005.

[32] A. Bergstrom, P. Pisani, V. Tenet, A. Wolk, and H. O. Adami. Overweight as an avoidable cause of cancer in europe. *Int J Cancer*, 91(3):421–30, 2001. Journal Article Meta-Analysis Research Support, Non-U.S. Gov't United States.

[33] E. E. Calle, C. Rodriguez, K. Walker-Thurmond, and M. J. Thun. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of u.s. adults. *N Engl J Med*, 348(17):1625–38, 2003.

[34] K. K. Carroll. Obesity as a risk factor for certain types of cancer. *Lipids*, 33(11):1055–9, 1998.

[35] S. C. Larsson and A. Wolk. Obesity and risk of non-hodgkin's lymphoma: a meta-analysis. *Int J Cancer*, 121(7):1564–70, 2007.

[36] S. C. Larsson and A. Wolk. Overweight, obesity and risk of liver cancer: a meta-analysis of cohort studies. *Br J Cancer*, 97(7):1005–8, 2007.

[37] C. M. Olsen, A. C. Green, D. C. Whiteman, S. Sadeghi, F. Kolahdooz, and P. M. Webb. Obesity and the risk of epithelial ovarian cancer: a systematic review and meta-analysis. *Eur J Cancer*, 43(4):690–709, 2007.

[38] A. Berrington de Gonzalez, S. Sweetland, and E. Spencer. A meta-analysis of obesity and the risk of pancreatic cancer. *Br J Cancer*, 89(3):519–23, 2003.

[39] S. J. Freedland and W. J. Aronson. Obesity and prostate cancer. *Urology*, 65(3):433–9, 2005.

[40] S. C. Larsson and A. Wolk. Obesity and the risk of gallbladder cancer: a meta-analysis. *Br J Cancer*, 96(9):1457–61, 2007. Journal Article Meta-Analysis Research Support, Non-U.S. Gov't England.

[41] S. C. Larsson and A. Wolk. Overweight and obesity and incidence of leukemia: A meta-analysis of cohort studies. *Int J Cancer*, 2007.

[42] G. Danaei, S. Vander Hoorn, A. D. Lopez, C. J. Murray, and M. Ezzati. Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. *Lancet*, 366(9499):1784–93, 2005.

[43] A McGuire. Written answer in response to parliamentary question by rht. hon. ruffley, d. mp. 7th november 2006. available at http://www.theyworkforyou.com/wrans/?id=2006-11-07c.90190.h. (accessed 12/12/2007)., 2006.

[44] A. M. Wolf and G. A. Colditz. The cost of obesity: the us perspective. *Pharmacoeconomics*, 5(Suppl 1):34–7, 1994.

[45] A. M. Wolf and G. A. Colditz. Current estimates of the economic cost of obesity in the united states. *Obes Res*, 6(2):97–106, 1998.

[46] Will Pay, Fred Kuchler, and Nicole Ballenger. The 2001 report the surgeon generals call to action to prevent and decrease overweight, 2001.

[47] J. Yates and C. Murphy. A cost benefit analysis of weight management strategies. *Asia Pac J Clin Nutr*, 15 Suppl:74–9, 2006.

[48] L. A. Tucker and G. M. Friedman. Obesity and absenteeism: an epidemiologic study of 10,825 employed adults. *Am J Health Promot*, 12(3):202–7, 1998.

[49] M. Moreau, F. Valente, R. Mak, E. Pelfrene, P. de Smet, G. De Backer, and M. Kornitzer. Obesity, body fat distribution and incidence of sick leave in the belgian workforce: the belstress study. *Int J Obes Relat Metab Disord*, 28(4):574–82, 2004.

[50] T Bungum, M Satterwhite, A W Jackson, and Jr. Morrow J. R. The relationship of body mass index, medical costs, and job absenteeism. *Am J Health Behav*, 27(4):456–462, 2003.

[51] R. P. Hertz, A. N. Unger, M. McDonald, M. B. Lustik, and J. Biddulph-Krentar. The impact of obesity on work limitations and cardiovascular risk factors in the u.s. workforce. *J Occup Environ Med*, 46(12):1196–203, 2004.

[52] A. J. Stunkard, M. S. Faith, and K. C. Allison. Depression and obesity. *Biol Psychiatry*, 54(3):330–7, 2003.

[53] R. L. Kolotkin, K. Meter, and G. R. Williams. Quality of life and obesity. *Obes Rev*, 2(4):219–29, 2001.

[54] J. Wardle, J. Waller, and M. J. Jarvis. Sex differences in the association of socioeconomic status with obesity. *Am J Public Health*, 92(8):1299–304, 2002.

[55] T. J. Parsons, C. Power, S. Logan, and C. D. Summerbell. Childhood predictors of adult obesity: a systematic review. *Int J Obes Relat Metab Disord*, 23 Suppl 8:S1–107, 1999.

[56] R Hardy, M Wadsworth, and D Kuh. The influence of childhood weight and socioeconomic status on change in adult body mass index in a British national birth cohort. *Int J Obes Relat Metab Disord*, 24(6):725–734, 2000.

[57] L. Ricciuto, V. Tarasuk, and A. Yatchew. Socio-demographic influences on food purchasing among canadian households. *Eur J Clin Nutr*, 60(6):778–90, 2006.

[58] L. E. Ricciuto and V. S. Tarasuk. An examination of income-related disparities in the nutritional quality of food selections among canadian households from 1986-2001. *Soc Sci Med*, 64(1):186–98, 2007.

[59] G. Turrell and A. M. Kavanagh. Socio-economic pathways to diet: modelling the association between socio-economic position and food purchasing behaviour. *Public Health Nutr*, 9(3):375–83, 2006.

[60] G. Turrell, T. Blakely, C. Patterson, and B. Oldenburg. A multilevel analysis of socioeconomic (small area) differences in household food purchasing behaviour. *J Epidemiol Community Health*, 58:208–215, Mar 2004.

[61] K Giskes, G Turrell, F J van Lenthe, J Brug, and J P Mackenbach. A multilevel study of socio-economic inequalities in food choice behaviour and dietary intake among the Dutch population: the GLOBE study. *Public Health Nutr*, 9(1):75–83, 2006.

[62] S. A. French, L. Harnack, and R. W. Jeffery. Fast food restaurant use among women in the pound of prevention study: dietary, behavioral and demographic correlates. *Int J Obes Relat Metab Disord*, 24(10):1353–9, 2000. French, S A Harnack, L Jeffery, R W Clinical Trial Randomized Controlled

Trial England International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity Int J Obes Relat Metab Disord. 2000 Oct;24(10):1353-9.

[63] S. C. Cummins, L. McKay, and S. MacIntyre. Mcdonald's restaurants and neighborhood deprivation in scotland and england. *Am J Prev Med*, 29(4):308–10, 2005.

[64] D. D. Reidpath, C. Burns, J. Garrard, M. Mahoney, and M. Townsend. An ecological study of the relationship between social and environmental determinants of obesity. *Health Place*, 8(2):141–5, 2002. Reidpath, Daniel D Burns, Cate Garrard, Jan Mahoney, Mary Townsend, Mardie England Health & place Health Place. 2002 Jun;8(2):141-5.

[65] K H Bellows-Riecken and R E Rhodes. A birth of inactivity? A review of physical activity and parenthood. *Prev Med*, 46(2):99–110, 2008.

[66] A Bauman, F Bull, T Chey, C L Craig, B E Ainsworth, J F Sallis, H R Bowles, M Hagstromer, M Sjostrom, and M Pratt. The International Prevalence Study on Physical Activity: results from 20 countries. *Int J Behav Nutr Phys Act*, 6(1):21, 2009.

[67] M. B. Livingstone, P. J. Robson, S. McCarthy, M. Kiely, K. Harrington, P. Browne, M. Galvin, N. J. Wareham, and K. L. Rennie. Physical activity patterns in a nationally representative sample of adults in ireland. *Public Health Nutr*, 4(5A):1107–16, 2001.

[68] S Macintyre and N Mutrie. Socio-economic differences in cardiovascular disease and physical activity: stereotypes and reality. *J R Soc Promot Health*, 124(2):66–69, 2004.

[69] F. Popham and R. Mitchell. Relation of employment status to socioeconomic position and physical activity types. *Prev Med*, 45(2-3):182–8, 2007.

[70] J. Panter, A. Jones, and M. Hillsdon. Equity of access to physical activity facilities in an english city. *Prev Med*, 46(4):303–7, 2008. Panter, Jenna Jones, Andy Hillsdon, Melvyn United States Preventive medicine Prev Med. 2008 Apr;46(4):303-7. Epub 2007 Nov 22.

[71] M Hillsdon, J Panter, C Foster, and A Jones. Equitable access to exercise facilities. *Am J Prev Med*, 32(6):506–508, 2007.

[72] S. Macintyre, L. Macdonald, and A. Ellaway. Do poorer people have poorer access to local resources and facilities? the distribution of local resources by area deprivation in glasgow, scotland. *Soc Sci Med*, 67(6):900–14, 2008. Macintyre, Sally Macdonald, Laura Ellaway, Anne Medical Research Council/United Kingdom Research Support, Non-U.S. Gov't England Social science & medicine (1982) Soc Sci Med. 2008 Sep;67(6):900-14. Epub 2008 Jul 1.

[73] S. Macintyre. Deprivation amplification revisited; or, is it always true that poorer places have poorer access to resources for healthy diets and physical activity? *Int J Behav Nutr Phys Act*, 4:32, 2007. Macintyre, Sally England The international journal of behavioral nutrition and physical activity Int J Behav Nutr Phys Act. 2007 Aug 7;4:32.

[74] W. Poortinga. The prevalence and clustering of four major lifestyle risk factors in an english adult population. *Prev Med*, 44(2):124–8, 2007.

[75] S Greenland. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–306, 2003.

[76] R. Murray, B. Frankowski, and H. Taras. Are soft drinks a scapegoat for childhood obesity? *J. Pediatr.*, 146:586–590, May 2005.

[77] C. O. Stubbs and A. J. Lee. The obesity epidemic: both energy intake and physical activity contribute. *Med. J. Aust.*, 181:489–491, Nov 2004.

[78] A. Morabia and M. C. Costanza. Does walking 15 minutes per day keep the obesity epidemic away? simulation of the efficacy of a populationwide campaign. *Am J Public Health*, 94(3):437–40, 2004.

[79] J. O. Hill, H. R. Wyatt, G. W. Reed, and J. C. Peters. Obesity and the environment: where do we go from here? *Science*, 299(5608):853–5, 2003.

[80] Boyd a Swinburn, Damien Jolley, Peter J Kremer, Arline D Salbe, and Eric Ravussin. Estimating the effects of energy imbalance on changes in body weight in children. *The American journal of clinical nutrition*, 83(4):859–63, April 2006.

[81] M. Wishnofsky. Caloric equivalents of gained or lost weight. *Am. J. Clin. Nutr.*, 6:542–546, 1958.

[82] K D Hall. What is the required energy deficit per unit weight loss? *International journal of obesity (2005)*, 32(3):573–6, March 2008.

[83] a Pietrobelli, D B Allison, S Heshka, M Heo, Z M Wang, a Bertkau, B Laferrère, M Rosenbaum, J F Aloia, F X Pi-Sunyer, and S B Heymsfield. Sexual dimorphism in the energy content of weight change. *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity*, 26(10):1339–48, October 2002.

[84] Edmund Christiansen, Lars Garby, and Thorkild I a Sø rensen. Quantitative analysis of the energy requirements for development of obesity. *Journal of theoretical biology*, 234(1):99–106, May 2005.

[85] K R Westerterp, J H Donkers, E W Fredrix, and P Boekhoudt. Energy intake, physical activity and body weight: a simulation model. *The British journal of nutrition*, 73(3):337–47, March 1995.

[86] D. T. Villareal, C. M. Apovian, R. F. Kushner, and S. Klein. Obesity in older adults: technical review and position statement of the American Society for Nutrition and NAASO, The Obesity Society. *Obes. Res.*, 13:1849–1863, Nov 2005.

[87] K. M. McTigue, R. Harris, B. Hemphill, L. Lux, S. Sutton, A. J. Bunton, and K. N. Lohr. Screening and interventions for obesity in adults: summary of the evidence for the U.S. Preventive Services Task Force. *Ann. Intern. Med.*, 139:933–949, Dec 2003.

[88] Marion J Franz, Jeffrey J VanWormer, a Lauren Crain, Jackie L Boucher, Trina Histon, William Caplan, Jill D Bowman, and Nicolas P Pronk. Weight-loss outcomes: a systematic review and meta-analysis of weight-loss clinical trials with a minimum 1-year follow-up. *Journal of the American Dietetic Association*, 107(10):1755–67, October 2007.

[89] a M Glenny, S O'Meara, a Melville, T a Sheldon, and C Wilson. The treatment and prevention of obesity: a systematic review of the literature. *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity*, 21(9):715–37, September 1997.

[90] K. Shaw, P. O'Rourke, C. Del Mar, and J. Kenardy. Psychological interventions for overweight or obesity. *Cochrane Database Syst Rev*, (2):CD003818, 2005.

[91] R. W. Jeffery and R. R. Wing. Long-term effects of interventions for weight loss using food provision and monetary incentives. *J Consult Clin Psychol*, 63(5):793–6, 1995.

[92] V.J. Stevens, E. Obarzanek, N.R. Cook, I. Lee, et al. Long-term weight loss and changes in blood pressure: results of the Trials of Hypertension Prevention, phase II. *Annals of Internal medicine*, 134(1):1, 2001.

[93] Rena R Wing and Suzanne Phelan. Long-term weight loss maintenance. *The American journal of clinical nutrition*, 82(1 Suppl):222S–225S, July 2005.

[94] S. J. Rodearmel, H. R. Wyatt, N. Stroebele, S. M. Smith, L. G. Ogden, and J. O. Hill. Small changes in dietary sugar and physical activity as an approach to preventing excessive weight gain: the america on the move family study. *Pediatrics*, 120(4):e869–79, 2007.

[95] J O Hill, J C Peters, and H R Wyatt. The role of public policy in treating the epidemic of global obesity. *Clinical pharmacology and therapeutics*, 81(5):772–5, May 2007.

[96] B Swinburn and G Egger. Preventive strategies against weight gain and obesity. *Obesity reviews : an official journal of the International Association for the Study of Obesity*, 3(4):289–301, November 2002.

[97] D. King. Foresight report on obesity. *Lancet*, 370(9601):1754; author reply 1755, 2007.

[98] Foresight. Tackling obesities: future choicesproject report. Oct 17, 2007., 2007.

[99] J. F. Sallis, A. Bauman, and M. Pratt. Environmental and policy interventions to promote physical activity. *Am J Prev Med*, 15(4):379–97, 1998.

[100] P. Hider. Environmental interventions to reduce energy intake or density. *A critical appraisal of the literature. New Zealand Health Technology Assessment Report*, 4(2), 2001.

[101] Marion Nestle. Food industry and health: mostly promises, little action. *Lancet*, 368(9535):564–5, August 2006.

[102] M. M. Mello, D. M. Studdert, and T. A. Brennan. Obesity–the new frontier of public health law. *N Engl J Med*, 354(24):2601–10, 2006.

[103] N.S. Lavery. Obesity: Stop all further researchand act. *BMJ: British Medical Journal*, 336(7634):7, 2008.

[104] James Fry and Willa Finley. The prevalence and costs of obesity in the EU. *Proceedings of the Nutrition Society*, 64(03):359–362, March 2007.

[105] A. Jack. Obesity plan lacks foresight. *Lancet*, 370(9598):1528–9, 2007.

[106] Kevin Murphy. An introduction to graphical models. Technical report, 2001.

[107] Dirk Husmeier. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, 19(17):2271–2282, 2003.

[108] Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1):89 – 109, 2001.

[109] I Gadaras and L Mikhailov. An interpretable fuzzy rule-based classification methodology for medical diagnosis. *Artif Intell Med*, 47(1):25–41, 2009.

[110] Bart Baesens, Michael Egmont-Petersen, Roberto Castelo, and Jan Vanthienen. Learning Bayesian network classifiers for credit scoring using Markov Chain Monte Carlo search. In *Proc. International Congress on Pattern Recognition*, pages 49–52, 2002.

[111] Z. Liu, S. Lin, and M. T. Tan. Sparse support vector machines with lp penalty for biomarker identification. *IEEE/ACM Trans Comput Biol Bioinform*, 7(1):100–7, 2010.

[112] K. S. Lynn, L. L. Li, Y. J. Lin, C. H. Wang, S. H. Sheng, J. H. Lin, W. Liao, W. L. Hsu, and W. H. Pan. A neural network model for constructing endophenotypes of common complex diseases: an application to male young-onset hypertension microarray data. *Bioinformatics*, 25(8):981–8, 2009.

[113] A. Simpson, V. Y. Tan, J. Winn, M. Svensen, C. M. Bishop, D. E. Heckerman, I. Buchan, and A. Custovic. Beyond atopy: multiple patterns of sensitization in relation to asthma in a birth cohort study. *Am J Respir Crit Care Med*, 181(11):1200–6, 2010.

[114] A. Li, J. Walling, S. Ahn, Y. Kotliarov, Q. Su, M. Quezado, J. C. Oberholtzer, J. Park, J. C. Zenklusen, and H. A. Fine. Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes. *Cancer Res*, 69(5):2091–9, 2009.

[115] A. D. Hummel, R. F. Maciel, R. G. Rodrigues, and I. T. Pisa. Application of artificial neural networks in renal transplantation: classification of nephrotoxicity and acute cellular rejection episodes. *Transplant Proc*, 42(2):471–2, 2010.

[116] G Caocci, R Baccoli, A Vacca, A Mastronuzzi, A Bertaina, E Piras, R Littera, F Locatelli, C Carcassi, and G La Nasa. Comparison between an artificial neural network and logistic regression in predicting acute graft-vs-host disease after unrelated donor hematopoietic stem cell transplantation in thalassemia patients. *Exp Hematol*, 38(5):426–433, 2010.

[117] J. Llorca, T. Dierssen-Sotos, I. Gomez-Acebo, A. Gonzalez-Castro, and E. Minambres. Artificial neural networks predict mortality after lung transplantation better than logistic regression. *J Heart Lung Transplant*, 28(11):1237–8, 2009.

[118] P. E. Puddu and A. Menotti. Artificial neural network versus multiple logistic function to predict 25-year coronary heart disease mortality in the seven countries study. *Eur J Cardiovasc Prev Rehabil*, 16(5):583–91, 2009.

[119] Y. N. Li, F. T. Luo, Y. M. Jiang, Y. R. Lu, J. L. Huang, and Z. B. Zhang. A prediction model of occupational manganese exposure based on artificial neural network. *Toxicol Mech Methods*, 19(5):337–45, 2009.

[120] J. Wang, M. Li, Y. T. Hu, and Y. Zhu. Comparison of hospital charge prediction models for gastric cancer patients: neural network vs. decision tree models. *BMC Health Serv Res*, 9:161, 2009.

[121] J. M. Quintana, A. Bilbao, A. Escobar, J. Azkarate, and J. I. Goenaga. Decision trees for indication of total hip replacement on patients with osteoarthritis. *Rheumatology (Oxford)*, 48(11):1402–9, 2009.

[122] A. M. Toschke, A. Beyerlein, and R. von Kries. Children at high risk for overweight: a classification and regression trees analysis approach. *Obes Res*, 13(7):1270–4, 2005.

[123] R. J. Marshall. The use of classification and regression trees in clinical epidemiology. *J Clin Epidemiol*, 54(6):603–9, 2001.

[124] N. Friedman and D. Koller. Being Bayesian about Bayesian network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50(1–2):95–125, 2003. Full version of UAI 2000 paper.

[125] Adriano V Werhli and Dirk Husmeier. Statistical Applications in Genetics and Molecular Biology Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Prior Knowledge Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

[126] Y. Xiao and M. R. Segal. Identification of yeast transcriptional regulation networks using multivariate random forests. *PLoS Comput Biol*, 5(6):e1000414, 2009.

[127] Junning Li, Z Jane Wang, Janice J Eng, and Martin J McKeown. Bayesian network modeling for discovering ”dependent synergies” among muscles in reaching movements. *IEEE transactions on bio-medical engineering*, 55(1):298–310, January 2008.

[128] Sinisa Pajevic and Dietmar Plenz. Efficient network reconstruction from dynamical cascades identifies small-world topology of neuronal avalanches. *PLoS computational biology*, 5(1):e1000271, January 2009.

[129] A Goodman and R Gatward. Who are we missing? Area deprivation and survey participation. *Eur J Epidemiol*, 23(6):379–387, 2008.

[130] National Centre for Social Research. *Lists of Variables and Derived Variables; Health Survey for England 2006.*, 2007.

[131] B. L. Heitmann and L. Lissner. Dietary underreporting by obese individuals–is it specific or non-specific? *BMJ*, 311:986–989, Oct 1995.

[132] K. L. Radimer and P. W. Harvey. Comparison of self-report of reduced fat and salt foods with sales and supply data. *Eur J Clin Nutr*, 52:380–382, May 1998.

[133] J. F. Sallis and B. E. Saelens. Assessment of physical activity by self-report: status, limitations, and future directions. *Res Q Exerc Sport*, 71:1–14, Jun 2000.

[134] L Basterfield, a J Adamson, K N Parkinson, U Maute, P X Li, and J J Reilly. Surveillance of physical activity in the UK is flawed: validation of the Health Survey for England Physical Activity Questionnaire. *Archives of disease in childhood*, 93(12):1054–8, December 2008.

[135] *Census 2001: General report for England and Wales.* Office for National Statistics, 2005. ISBN 1403987688.

[136] Paul Boyle and Danny Dorling. Guest editorial: the 2001 UK census: remarkable resource or bygone legacy of the 'pencil and paper era'? *Area*, 36(2):101–110, 2004.

[137] Youngjoo Lee, Namkug Kim, Kyoung-Sik Cho, Suk-Ho Kang, Dae Y Kim, Yoon Y Jung, and Jeong K Kim. Bayesian Classifier for Predicting Malignant Renal Cysts on MDCT: Early Clinical Experience. *Am. J. Roentgenol.*, 193(2):W106—-111, August 2009.

[138] Martin Neil and Norman Fenton. Using bayesian networks to model the operational risk to information technology infrastructure in financial institutions. *Journal of Financial Transformation*, 22:131–138, 2008.

[139] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann, September 1988.

[140] David Heckerman. A tutorial on learning with bayesian networks. In *Learning in graphical models*, pages 301–354. MIT Press, 1999.

[141] Colin Howson and Peter Urbach. *Scientific Reasoning: The Bayesian Approach*. Open Court, third edition, April 2005.

[142] D. Heckerman, D. Geiger, and D. M. Chickering. Learning bayesian networks - the combination of knowledge and statistical-data. *Machine Learning*, 20(3):197–243, 1995.

[143] A. P. Grzegorczyk, J. L. Weyher, P. R. Hageman, and P. K. Larsen. Influence of sapphire annealing in a trimethylaluminum atmosphere on gan epitaxy by metal-organic chemical vapor deposition. *Thin Solid Films*, 516(8):2314–2320, 2008. 272PQ Times Cited:0 Cited References Count:30.

[144] D. Madigan, J. York, and D. Allard. Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique*, pages 215–232, 1995.

[145] W Buntine. Theory refinement on Bayesian networks. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, volume 91, pages 52–60. Citeseer, 1991.

[146] D. J. Spiegelhalter and S. L. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20(5):579–605, 1990.

[147] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–710, 1995. With discussion and a rejoinder by the author.

[148] R.W. Robinson. Counting labeled acyclic digraphs. In Frank Harary, editor, *New Directions in the Theory of Graphs*, pages 239–273. Academic Press, New York, 1973.

[149] D Madigan, A Raftery, J York, J Bradshaw, and R Almond. Strategies for graphical model selection, 1993.

[150] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[151] R Castelo and T Kocka. On inclusion-driven learning of Bayesian networks. *The Journal of Machine Learning Research*, 4:574, 2003.

[152] D M Chickering. Learning equivalence classes of Bayesian-network structures. *The Journal of Machine Learning Research*, 2:445–498, 2002.

[153] B Ellis and W Wong. Sampling Bayesian Networks Quickly. In *Interface conference, Pasadena CA.*, 2006.

[154] Mikko Koivisto. Advances in exact bayesian structure discovery in bayesian networks. In *Proceedings of the Proceedings of the Twenty-Second Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 241–248, Arlington, Virginia, 2006. AUAI Press.

[155] Mikko Koivisto and Kismat Sood. Exact Bayesian Structure Discovery in Bayesian Networks. *J. of Machine Learning Research*, 5:549–573, 2004.

[156] Daniel Eaton and Kevin Murphy. Exact Bayesian structure learning from uncertain interventions. In *AI & Statistics*, 2007.

[157] Michael D Linderman, Robert Bruggner, Vivek Athalye, Teresa H Meng, Narges Bani Asadi, and Garry P Nolan. High-throughput Bayesian network learning using heterogeneous multicore computers. In *ICS '10: Proceedings of the 24th ACM International Conference on Supercomputing*, pages 95–104, New York, NY, USA, 2010. ACM.

[158] Narges Bani Asadi, Teresa H Meng, and Wing H Wong. Reconfigurable computing for learning Bayesian networks. In *FPGA '08: Proceedings of the 16th international ACM/SIGDA symposium on Field programmable gate arrays*, pages 203–211, New York, NY, USA, 2008. ACM.

[159] Faming Liang, Chuanhai Liu, and Raymond J. Carroll. Stochastic approximation in monte carlo computation. *Journal of the American Statistical Association*, 102:305–320, 2007.

[160] Faming Liang and Jian Zhang. Learning bayesian networks for discrete data. *Comput. Stat. Data Anal.*, 53(4):865–876, 2009.

[161] Stephen P. Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455, 1998.

[162] J. Venna, S. Kaski, and J. Peltonen. Visualizations for assessing convergence and mixing of MCMC. *Machine Learning: ECML 2003*, pages 432–443, 2003.

[163] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[164] S Kirkpatrick, C D Gelatt, and M P Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, 1983.

[165] K. K. Davison, L. A. Francis, and L. L. Birch. Links between parents' and girls' television viewing behaviors: a longitudinal examination. *J Pediatr*, 147(4):436–42, 2005.

[166] K. K. Davison and L. L. Birch. Obesigenic families: parents' physical activity and dietary intake patterns predict girls' risk of overweight. *Int J Obes Relat Metab Disord*, 26(9):1186–93, 2002. Davison, K Krahnstoever Birch, L Lipps R01 HD 32973/HD/NICHD NIH HHS/United States R01 HD032973-06/HD/NICHD NIH HHS/United States Research Support, U.S. Gov't, P.H.S. England International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity Nihms61922 Int J Obes Relat Metab Disord. 2002 Sep;26(9):1186-93.

[167] J. K. Pitner. Obesity in the elderly. *Consult Pharm*, 20:498–513, Jun 2005.

[168] G Cooper, D Heckerman, and C Meek. A Bayesian approach to causal discovery. Technical report, Technical report, Microsoft Research Advanced Technology Division, Microsoft Corporation, Technical Report MSR-TR-97-05, 1997, 1997.

[169] A. Amuzu, C. Carson, H. C. Watt, D. A. Lawlor, and S. Ebrahim. Influence of area and individual lifecourse deprivation on health behaviours: findings from the british women's heart and health study. *Eur J Cardiovasc Prev Rehabil*, 16(2):169–73, 2009. Amuzu, Antoinette Carson, Claire Watt, Hilary C Lawlor, Debbie A Ebrahim, Shah British Heart Foundation/United Kingdom Department of Health/United Kingdom Comparative Study Research Support, Non-U.S. Gov't England European journal of cardiovascular prevention and rehabilitation : official journal of the European Society of Cardiology, Working Groups on Epidemiology & Prevention and Cardiac Rehabilitation and Exercise Physiology Eur J Cardiovasc Prev Rehabil. 2009 Apr;16(2):169-73.

[170] E. Stamatakis. *Risks factors for cardiovascular disease: Physical Activity*, volume 2, chapter 5, pages 107–142. National Statistics, 2004.

[171] D. Rucker, R. Padwal, S. K. Li, C. Curioni, and D. C. Lau. Long term pharmacotherapy for obesity and overweight: updated meta-analysis. *Bmj*, 335(7631):1194–9, 2007.

[172] C. D. Sjostrom, L. Lissner, H. Wedel, and L. Sjostrom. Reduction in incidence of diabetes, hypertension and lipid disturbances after intentional weight loss induced by bariatric surgery: the SOS Intervention Study. *Obes. Res.*, 7:477–484, Sep 1999.

[173] W. Poortinga. Do health behaviors mediate the association between social capital and health? *Preventive medicine*, 43(6):488–493, 2006.

[174] S Allender, C Foster, and A Boxer. Occupational and nonoccupational physical activity and the social determinants of physical activity: results from the Health Survey for England. *J Phys Act Health*, 5(1):104–116, 2008.

[175] C. Foster, M. Hillsdon, A. Jones, C. Grundy, P. Wilkinson, M. White, B. Sheehan, N. Wareham, and M. Thorogood. Objective measures of the environment and physical activity–results of the environment and physical activity study in english adults. *J Phys Act Health*, 6 Suppl 1:S70–80, 2009.

[176] A. Jones, M. Hillsdon, and E. Coombes. Greenspace access, use, and physical activity: understanding the effects of area deprivation. *Prev Med*, 49(6):500–5, 2009. Jones, Andy Hillsdon, Melvyn Coombes, Emma Research Support, Non-U.S. Gov't United States Preventive medicine Prev Med. 2009 Dec;49(6):500-5. Epub 2009 Oct 24.

[177] T. J. Parsons, C. Thomas, and C. Power. Estimated activity patterns in british 45 year olds: cross-sectional findings from the 1958 british birth cohort. *Eur J Clin Nutr*, 63(8):978–85, 2009.

[178] T. Gorely, S. Biddle, S. Marshall, N. Cameron, and L. Cassey. The association between distance to school, physical activity and sedentary behaviors in adolescents: project stil. *Pediatr Exerc Sci*, 21(4):450–61, 2009. Gorely, Trish Biddle, Stuart Marshall, Simon Cameron, Noel Cassey, Louise British Heart Foundation/United Kingdom Multicenter Study Research Support, Non-U.S. Gov't United States Pediatric exercise science Pediatr Exerc Sci. 2009 Nov;21(4):450-61.

[179] A. A. Lake, T. Townshend, S. Alvanides, E. Stamp, and A. J. Adamson. Diet, physical activity, sedentary behaviour and perceptions of the environment in young adults. *J Hum Nutr Diet*, 22(5):444–54, 2009. Lake, A A Townshend, T Alvanides, S Stamp, E Adamson, A J Research Support,

N.I.H., Extramural England Journal of human nutrition and dietetics : the official journal of the British Dietetic Association J Hum Nutr Diet. 2009 Oct;22(5):444-54.

[180] S. A. McLure, C. D. Summerbell, and J. J. Reilly. Objectively measured habitual physical activity in a highly obesogenic environment. *Child Care Health Dev*, 35(3):369–75, 2009.

[181] P. Tucker, J. D. Irwin, J. Gilliland, M. He, K. Larsen, and P. Hess. Environmental influences on physical activity levels in youth. *Health Place*, 15(1):357–63, 2009. Tucker, Patricia Irwin, Jennifer D Gilliland, Jason He, Meizi Larsen, Kristian Hess, Paul Research Support, Non-U.S. Gov't England Health & place Health Place. 2009 Mar;15(1):357-63. Epub 2008 Jul 9.

[182] National diet nutrition survey: headline results from year 1 (2008/2009). `http://www.food.gov.uk/science/dietarysurveys/ndnsdocuments/ndns0809year1`, August 2010.

[183] E. Stamatakis, U. Ekelund, and N. J. Wareham. Temporal trends in physical activity in england: the health survey for england 1991 to 2004. *Prev Med*, 45(6):416–23, 2007.

[184] E. Stamatakis and M. Chaudhury. Temporal trends in adults' sports participation patterns in england between 1997 and 2006: the health survey for england. *Br J Sports Med*, 42(11):601–8, 2008.

[185] Murray Turoff and Harold Linstone. The delphi method: Techniques and applications. 18(3):363, 2002.

[186] R. Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. online, 2005.

[187] Nuoo-Ting Molitor, Nicky Best, Chris Jackson, and Sylvia Richardson. Using bayesian graphical models to model biases in observational studies and to combine multiple sources of data: application to low birth weight and water disinfection by-products. *Journal Of The Royal Statistical Society Series A*, 172(3):615–637, 2009.

[188] A. Romero-Corral, V. K. Somers, J. Sierra-Johnson, R. J. Thomas, M. L. Collazo-Clavell, J. Korinek, T. G. Allison, J. A. Batsis, F. H. Sert-Kuniyoshi, and F. Lopez-Jimenez. Accuracy of body mass index in diagnosing obesity in the adult general population. *Int J Obes*, 32(6):959–966, 2008.

[189] M. D. Jensen. Role of body fat distribution and the metabolic complications of obesity. *J Clin Endocrinol Metab*, 93(11 Suppl 1):S57–63, 2008.

[190] J. P. Reis, C. A. Macera, M. R. Araneta, S. P. Lindsay, S. J. Marshall, and D. L. Wingard. Comparison of overall obesity and body fat distribution in predicting risk of mortality. *Obesity (Silver Spring)*, 17(6):1232–9, 2009.

[191] D Canoy. Distribution of body fat and risk of coronary heart disease in men and women. *Curr Opin Cardiol*, 23(6):591–598, 2008.

[192] J Stevens, E G Katz, and R R Huxley. Associations between gender, age and waist circumference. *Eur J Clin Nutr*, 64(1):6–15, 2010.

[193] J L Kuk, S Lee, S B Heymsfield, and R Ross. Waist circumference and abdominal adipose tissue distribution: influence of age and sex. *Am J Clin Nutr*, 81(6):1330–1334, 2005.

[194] D R Wagner and V H Heyward. Measures of body composition in blacks and whites: a comparative review. *Am J Clin Nutr*, 71(6):1392–1402, 2000.

[195] S. B. Sisson, P. T. Katzmarzyk, S. R. Srinivasan, W. Chen, D. S. Freedman, C. Bouchard, and G. S. Berenson. Ethnic differences in subcutaneous adiposity and waist girth in children and adolescents. *Obesity (Silver Spring)*, 17(11):2075–81, 2009. Sisson, Susan B Katzmarzyk, Peter T Srinivasan, Sathanur R Chen, Wei Freedman, David S Bouchard, Claude Berenson, Gerald S HD043820/HD/NICHD NIH HHS/United States HL38844/HL/NHLBI NIH HHS/United States Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't United States Obesity (Silver Spring, Md.) Nihms140759 Obesity (Silver Spring). 2009 Nov;17(11):2075-81. Epub 2009 Apr 23.

[196] J F Carroll, A L Chiapa, M Rodriquez, D R Phelps, K M Cardarelli, J K Vishwanatha, S Bae, and R Cardarelli. Visceral fat, waist circumference, and BMI: impact of race/ethnicity. *Obesity (Silver Spring)*, 16(3):600–607, 2008.

[197] A. Must, L. G. Bandini, D. J. Tybor, I. Janssen, R. Ross, and W. H. Dietz. Behavioral risk factors in relation to visceral adipose tissue deposition in adolescent females. *Int J Pediatr Obes*, 3 Suppl 1:28–36, 2008.

[198] E. A. Molenaar, J. M. Massaro, P. F. Jacques, K. M. Pou, R. C. Ellison, U. Hoffmann, K. Pencina, S. D. Shadwick, R. S. Vasan, C. J. O'Donnell, and C. S. Fox. Association of lifestyle factors with abdominal subcutaneous and visceral adiposity: the framingham heart study. *Diabetes Care*, 32(3):505–10, 2009.

[199] D Canoy, R Luben, A Welch, S Bingham, N Wareham, N Day, and K T Khaw. Fat distribution, body mass index and blood pressure in 22,090 men and women in the Norfolk cohort of the European Prospective Investigation into Cancer and Nutrition (EPIC-Norfolk) study. *J Hypertens*, 22(11):2067–2074, 2004.

[200] M Akbartabartoori, M E Lean, and C R Hankey. Relationships between cigarette smoking, body size and body shape. *Int J Obes (Lond)*, 29(2):236–243, 2005.

[201] C. Pisinger, U. Toft, and T. Jorgensen. Can lifestyle factors explain why body mass index and waist-to-hip ratio increase with increasing tobacco consumption? the inter99 study. *Public Health*, 123(2):110–5, 2009.

[202] D. L. Reas, J. F. Nygard, and T. Sorensen. Do quitters have anything to lose? changes in body mass index for daily, never, and former smokers over an 11-year period (1990–2001). *Scand J Public Health*, 37(7):774–7, 2009.

[203] J. S. Tolstrup, J. Halkjaer, B. L. Heitmann, A. M. Tjonneland, K. Overvad, T. I. Sorensen, and M. N. Gronbaek. Alcohol drinking frequency in relation to subsequent changes in waist circumference. *Am J Clin Nutr*, 87(4):957–63, 2008.

[204] M. Sowers, H. Zheng, K. Tomey, C. Karvonen-Gutierrez, M. Jannausch, X. Li, M. Yosef, and J. Symons. Changes in body composition in women over six years at midlife: ovarian and chronological aging. *J Clin Endocrinol Metab*, 92(3):895–901, 2007.

[205] R Ross, J Rissanen, H Pedwell, J Clifford, and P Shragge. Influence of diet and exercise on skeletal muscle and visceral adipose tissue in men. *J Appl Physiol*, 81(6):2445–2455, 1996.

[206] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

[207] S. Derksen and H. J. Keselman. Backward, forward and stepwise automated subset selection algorithms : frequency of obtaining authentic and noise variables. *British journal of mathematical & statistical psychology*, 45(2):265–282, 1992.

[208] Y. K. Tu, M. Kellett, V. Clerehugh, and M. S. Gilthorpe. Problems of correlations between explanatory variables in multiple regression analyses in the dental literature. *Br Dent J*, 199(7):457–461, 2005. 10.1038/sj.bdj.4812743.

[209] J. Miles and M. Shevlin. *Applying regression & correlation: A guide for students and researchers*. Sage Publications Ltd, 2001.

[210] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 2006. corr. 2nd printing edition, October 2007.

[211] J. Pinilla, B. Gonzalez, P. Barber, and Y. Santana. Smoking in young adolescents: an approach with multilevel discrete choice models. *J Epidemiol Community Health*, 56:227–232, Mar 2002.

[212] Scott Davies and Andrew Moore. Mixnets: Learning bayesian networks with mixtures of discrete and continuous attributes. 2000.

[213] Kevin P. Murphy. A variational approximation for bayesian networks with discrete and continuous latent variables. pages 457–466, 1999.

[214] I. T. Jolliffe. *Principal component analysis*. Springer, 2002.

[215] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. Numerical recipes 3rd edition: The art of scientific computing. 2007.

[216] Edwin B Wilson. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[217] The food labelling regulations (uk), 1996.

[218] FDF. Response from the food and drink federation (fdf) to: Dg sanco consultative document labelling: Competitiveness, consumer information and better regulation for the eu, 2006.

[219] T. C. Beard, C. A. Nowson, and M. D. Riley. Traffic-light food labels. *Med J Aust*, 186(1):19, 2007.

[220] Food standards agency. fsa news. september 2006., 2006.

[221] Jonathan L Blitstein and W Douglas Evans. Use of nutrition facts panels among adults who make household food purchasing decisions. *Journal of nutrition education and behavior*, 38(6):360–4, 2003.

[222] E. van Kleef, H. van Trijp, F. Paeps, and L. Fernandez-Celemin. Consumer preferences for front-of-pack calories labelling. *Public Health Nutr*, 11:203–213, Feb 2008.

[223] Jason Switt. Labeling around the globe: helping to direct food flow. *Journal of the American Dietetic Association*, 107(2):199–200, February 2007.

[224] A. M. Prentice and S. A. Jebb. Fast foods, energy density and obesity: a possible mechanistic link. *Obes Rev*, 4(4):187–94, 2003.

[225] S.J. Nielsen and B.M. Popkin. Patterns and trends in food portion sizes, 1977-1998. *JAMA: the journal of the American Medical Association*, 289(4):450, 2003.

[226] S.A. Bowman and B.T. Vinyard. Fast food consumption of US adults: impact on energy and nutrient intakes and overweight status. *Journal of the American College of Nutrition*, 23(2):163, 2004.

[227] M. Nestle. Food marketing and childhood obesity–a matter of policy. *N Engl J Med*, 354(24):2527–9, 2006.

[228] S. Burton, E. H. Creyer, J. Kees, and K. Huggins. Attacking the obesity epidemic: the potential health benefits of providing nutrition information in restaurants. *Am J Public Health*, 96(9):1669–75, 2006.

[229] J. Backstrand, M. G. Wootan, L.R. Young, and J. Hurley. *Fat chance, Center for Science in the Public Interest.* Washington DC., 1997.

[230] M. G. Wootan. Need for and effectiveness of menu labeling. *J Am Diet Assoc*, 107(1):33–4; author reply 34–5, 2007.

[231] M. G. Wootan and M. Osborn. Availability of nutrition information from chain restaurants in the united states. *Am J Prev Med*, 30(3):266–8, 2006.

[232] M. G. Wootan, M. Osborn, and C. J. Malloy. Availability of point-of-purchase nutrition information at a fast-food restaurant. *Prev Med*, 43(6):458–9, 2006.

[233] Rebecca a Krukowski, Jean Harvey-Berino, Jane Kolodinsky, Rashmi T Narsana, and Thomas P Desisto. Consumers may not use or understand calorie labeling in restaurants. *Journal of the American Dietetic Association*, 106(6):917–20, June 2006.

[234] T. V. Kral, L. S. Roe, and B. J. Rolls. Does nutrition information about the energy density of meals affect food intake in normal-weight women? *Appetite*, 39(2):137–45, 2002.

[235] M.T. Conklin, C.U. Lambert, and D.A. Cranage. Nutrition information at point of selection could benefit college students. *Topics in Clinical Nutrition*, 20(2):90, 2005.

[236] J.C. Kozup, E.H. Creyer, and S. Burton. Making healthful food choices: The influence of health claims and nutrition information on consumers evaluations of packaged food products and restaurant menu items. *Journal of Marketing*, 67(2):19–34, 2003.

[237] Dara Bergen and Ming-Chin Yeh. Effects of energy-content labels and motivational posters on sales of sugar-sweetened beverages: stimulating sales of diet drinks among adults study. *Journal of the American Dietetic Association*, 106(11):1866–9, November 2006.

[238] J.D. Seymour, A. Lazarus Yaroch, M. Serdula, H.M. Blanck, and L.K. Khan. Impact of nutrition environmental interventions on point-of-purchase behavior in adults: a review. *Preventive Medicine*, 39:108–136, 2004.

[239] A. Fiske and K.W. Cullen. Effects of promotional materials on vending sales of low-fat items in teachers' lounges. *Journal of the American Dietetic Association*, 104(1):90–93, 2004.

[240] S.A. French, R.W. Jeffery, M. Story, K.K. Breitlow, J.S. Baxter, P. Hannan, and M.P. Snyder. Pricing and promotion effects on low-fat vending snack purchases: the CHIPS Study. *American Journal of Public Health*, 91(1):112, 2001.

[241] R.A. Krukowski, J. Harvey-Berino, J. Kolodinsky, R.T. Narsana, and T.P. DeSisto. Consumers may not use or understand calorie labeling in restaurants. *Journal of the American Dietetic Association*, 106(6):917–920, 2006.

[242] Beth Antonuk and Lauren G Block. The effect of single serving versus entire package nutritional information on consumption norms and actual consumption of a snack food. *Journal of nutrition education and behavior*, 38(6):365–70, 2006.

[243] E. Finkelstein, S. French, J. N. Variyam, and P. S. Haines. Pros and cons of proposed interventions to promote healthy eating. *Am J Prev Med*, 27(3 Suppl):163–71, 2004.

[244] K.D. Brownell, A.J. Stunkard, and J.M. Albaum. Evaluation and modification of exercise patterns in the natural environment. *American Journal of Psychiatry*, 137(12):1540, 1980.

[245] A. Blamey, N. Mutrie, and A. Tom. Health promotion by encouraged use of stairs. *Bmj*, 311(7000):289, 1995.

[246] R.E. Andersen, S.C. Franckowiak, J. Snyder, S.J. Bartlett, and K.R. Fontaine. Can inexpensive signs encourage the use of stairs? Results from a community intervention. *Annals of Internal Medicine*, 129(5):363, 1998.

[247] WD Russell, DA Dzewaltowski, and GJ Ryan. The effectiveness of a point-of-decision prompt in deterring sedentary behavior. *American journal of health promotion: AJHP*, 13(5):257, 1999.

[248] M. A. Flynn, D. A. McNeil, B. Maloff, D. Mutasingwa, M. Wu, C. Ford, and S. C. Tough. Reducing obesity and related chronic disease risk in children and youth: a synthesis of evidence with 'best practice' recommendations. *Obes Rev*, 7 Suppl 1:7–66, 2006.

[249] M. Nestle and M. F. Jacobson. Halting the obesity epidemic: a public health policy approach. *Public Health Rep*, 115(1):12–24, 2000. Journal Article Research Support, Non-U.S. Gov't United states 1974).

[250] E.B. Kahn, L.T. Ramsey, R.C. Brownson, G.W. Heath, E.H. Howze, K.E. Powell, E.J. Stone, M.W. Rajab, and P. Corso. The effectiveness of interventions to increase physical activity. *American journal of preventive medicine*, 22(4S):73–107, 2002.

[251] White house proceedings: White house conference on food nutrition and health. U.S. Government Printing Office Washington, DC., 1970.

[252] Cécile Knai, Joceline Pomerleau, Karen Lock, and Martin McKee. Getting children to eat more fruit and vegetables: a systematic review. *Preventive medicine*, 42(2):85–95, February 2006.

[253] T. Baranowski, M. Davis, K. Resnicow, J. Baranowski, C. Doyle, L. S. Lin, M. Smith, and D. T. Wang. Gimme 5 fruit, juice, and vegetables for fun and health: outcome evaluation. *Health Educ Behav*, 27(1):96–111, 2000.

[254] F. Wong, M. Huhman, L. Asbury, R. Bretthauer-Mueller, S. McCarthy, P. Londe, and C. Heitzler. VERBa social marketing campaign to increase physical activity among youth. *Preventing Chronic Disease*, 1(3), 2004.

[255] ML Hunter, IG Chestnutt, SM Evans, and AC Withecombe. Fluid for thought: availability of drinks in primary and secondary schools in Cardiff, UK. *International Journal of Paediatric Dentistry*, 14(4):267–271, 2004.

[256] L.D. Johnston, J. Delva, and P.M. O'Malley. Soft drink availability, contracts, and revenues in American secondary schools. *American journal of preventive medicine*, 33(4):S209–S225, 2007.

[257] J.M. McGinnis, J.A. Gootman, and V.I. Kraak. *Food marketing to children and youth: threat or opportunity?* Natl Academy Pr, 2006.

[258] G.A. Bray, S.J. Nielsen, and B.M. Popkin. Consumption of high-fructose corn syrup in beverages may play a role in the epidemic of obesity. *The American journal of clinical nutrition*, 79(4):537, 2004.

[259] L. Harnack, J. Stang, and M. Story. Soft Drink Consumption Among US Children and Adolescents:: Nutritional Consequences. *Journal of the American Dietetic Association*, 99(4):436–441, 1999.

[260] J.L. Wiecha, D. Finkelstein, P.J. Troped, M. Fragala, and K.E. Peterson. School vending machine use and fast-food restaurant use are associated with

sugar-sweetened beverage intake in youth. *Journal of the American Dietetic Association*, 106(10):1624–1630, 2006.

[261] Cara B Ebbeling, Henry a Feldman, Stavroula K Osganian, Virginia R Chomitz, Sheila J Ellenbogen, and David S Ludwig. Effects of decreasing sugar-sweetened beverage consumption on body weight in adolescents: a randomized, controlled pilot study. *Pediatrics*, 117(3):673–80, March 2006.

[262] Nutritional standards for school lunches and other school food. school food trust., 2006.

[263] M. Nelson, K. Lowes, and V. Hwang. The contribution of school meals to food consumption and nutrient intakes of young people aged 4-18 years in england. *Public Health Nutr*, 10(7):652–62, 2007. members of the Nutrition Group, School Meals Review Panel, Department for Education and Skills.

[264] R. Gould, J. Russell, and M. E. Barker. School lunch menus and 11 to 12 year old children's food choice in three secondary schools in england-are the nutritional standards being met? *Appetite*, 46(1):86–92, 2006.

[265] J. Kolodinsky, J.R. Harvey-Berino, L. Berlin, R.K. Johnson, and T.W. Reynolds. Knowledge of current dietary guidelines and food choice by college students: better eaters have higher knowledge of dietary guidance. *Journal of the American Dietetic Association*, 107(8):1409–1413, 2007.

[266] IS Rogers, AR Ness, K. Hebditch, LR Jones, and PM Emmett. Quality of food eaten in English primary schools: school dinners vs packed lunches. *European journal of clinical nutrition*, 61(7):856–864, 2007.

[267] W. P. James, M. Nelson, A. Ralph, and S. Leather. Socioeconomic determinants of health. the contribution of nutrition to inequalities in health. *Bmj*, 314(7093):1545–9, 1997.

[268] D; Acheson. Independent inquiry into inequalities in health. london: Stationery office. 1998.

[269] T. Pearson, J. Russell, M. J. Campbell, and M. E. Barker. Do 'food deserts' influence fruit and vegetable consumption?–a cross-sectional study. *Appetite*, 45(2):195–7, 2005.

[270] S. Cummins and S. Macintyre. ”food deserts”–evidence and assumption in health policy making. *Bmj*, 325(7361):436–8, 2002.

[271] J. Pearce, T. Blakely, K. Witten, and P. Bartie. Neighborhood deprivation and access to fast-food retailing: a national study. *Am J Prev Med*, 32(5):375–82, 2007.

[272] L. Macdonald, S. Cummins, and S. Macintyre. Neighbourhood fast food environment and area deprivation-substitution or concentration? *Appetite*, 49(1):251–4, 2007.

[273] M. Ashe, D. Jernigan, R. Kline, and R. Galaz. Land use planning and the control of alcohol, tobacco, firearms, and fast food restaurants. *Am J Public Health*, 93(9):1404–8, 2003.

[274] S. Okie. New york to trans fats: you’re out! *N Engl J Med*, 356(20):2017–21, 2007.

[275] D. Mozaffarian, M. B. Katan, A. Ascherio, M. J. Stampfer, and W. C. Willett. Trans fatty acids and cardiovascular disease. *N Engl J Med*, 354(15):1601–13, 2006.

[276] E. Ravussin, S. Lillioja, W. C. Knowler, L. Christin, D. Freymond, W. G. Abbott, V. Boyce, B. V. Howard, and C. Bogardus. Reduced rate of energy expenditure as a risk factor for body-weight gain. *N Engl J Med*, 318(8):467–72, 1988.

[277] M. A. Papas, A. J. Alberg, R. Ewing, K. J. Helzlsouer, T. L. Gary, and A. C. Klassen. The built environment and obesity. *Epidemiol Rev*, 29:129–43, 2007.

[278] E. Moore, B. A. Richter, C. K. Patton, and S. A. Lear. Mapping stairwell accessibility in vancouver’s downtown core. *Can J Public Health*, 97(2):118–20, 2006.

[279] R. Pendola and S. Gen. Bmi, auto use, and the urban environment in san francisco. *Health Place*, 13(2):551–6, 2007.

[280] J. F. Sallis and K. Glanz. The role of built environments in physical activity, eating, and obesity in childhood. *Future Child*, 16(1):89–108, 2006.

[281] C. L. Hayne, P. A. Moran, and M. M. Ford. Regulating environments to reduce obesity. *J Public Health Policy*, 25(3-4):391–407, 2004.

[282] L. D. Frank, M. A. Andresen, and T. L. Schmid. Obesity relationships with community design, physical activity, and time spent in cars. *Am J Prev Med*, 27(2):87–96, 2004.

[283] J. Wakefield. Fighting obesity through the built environment. *Environ Health Perspect*, 112(11):A616–8, 2004.

[284] R. J. Stokes, J. MacDonald, and G. Ridgeway. Estimating the effects of light rail transit on health care costs. *Health Place*, 14(1):45–58, 2008.

[285] X. Guo, B. M. Popkin, T. A. Mroz, and F. Zhai. Food price policy can favorably alter macronutrient intake in china. *J Nutr*, 129(5):994–1001, 1999.

[286] T. Marshall. Exploring a fiscal food policy: the case of diet and ischaemic heart disease. *Bmj*, 320(7230):301–5, 2000.

[287] Jr. Garson, A. and C. L. Engelhard. Attacking obesity: lessons from smoking. *J Am Coll Cardiol*, 49(16):1673–5, 2007.

[288] G. E. Guindon, S. Tobin, and D. Yach. Trends and affordability of cigarette prices: ample room for tax increases and related health gains. *Tob Control*, 11(1):35–43, 2002.

[289] M. F. Jacobson and K. D. Brownell. Small taxes on soft drinks and snack foods to promote health. *Am J Public Health*, 90(6):854–7, 2000.

[290] D. Kim and I. Kawachi. Food taxation and pricing strategies to ”thin out” the obesity epidemic. *Am J Prev Med*, 30(5):430–7, 2006.

[291] F; Kuchler, A; Tegene, and Harris JM;. Taxing snack foods: manipulating diet quality or financing information programs? *Rev Agric Econ*, (27):4–20, 2004.

[292] B. C. Tohill, J. Seymour, M. Serdula, L. Kettel-Khan, and B. J. Rolls. What epidemiologic studies tell us about the relationship between fruit and vegetable consumption and body weight. *Nutr Rev*, 62(10):365–74, 2004.

[293] L. A. Bazzano. The high cost of not consuming fruits and vegetables. *J Am Diet Assoc*, 106(9):1364–8, 2006.

[294] A. D. Liese, K. E. Weis, D. Pluto, E. Smith, and A. Lawson. Food store types, availability, and cost of foods in a rural environment. *J Am Diet Assoc*, 107(11):1916–23, 2007.

[295] K. M. Jetter and D. L. Cassady. The availability and cost of healthier food alternatives. *Am J Prev Med*, 30(1):38–44, 2006.

[296] D. Cassady, K. M. Jetter, and J. Culp. Is price a barrier to eating more fruits and vegetables for low-income families? *J Am Diet Assoc*, 107(11):1909–15, 2007.

[297] D. R. Herman, G. G. Harrison, and E. Jenks. Choices made by low-income women provided with an economic supplement for fresh fruit and vegetable purchase. *J Am Diet Assoc*, 106(5):740–4, 2006.

[298] D. R. Herman, G. G. Harrison, A. A. Afifi, and E. Jenks. Effect of a targeted subsidy on intake of fruits and vegetables among low-income women in the special supplemental nutrition program for women, infants, and children. *Am J Public Health*, 98(1):98–105, 2008.

[299] S. A. French. Public health strategies for dietary change: schools and workplaces. *J Nutr*, 135(4):910–2, 2005.

[300] N. Lien, L. A. Lytle, and K. I. Klepp. Stability in consumption of fruit, vegetables, and sugary foods in a cohort from age 14 to age 21. *Prev Med*, 33(3):217–26, 2001.

[301] W. H. Dietz and S. L. Gortmaker. Preventing obesity in children and adolescents. *Annu Rev Public Health*, 22:337–53, 2001.

[302] P. Hannan, S. A. French, M. Story, and J. A. Fulkerson. A pricing strategy to promote sales of lower fat foods in high school cafeterias: acceptability and sensitivity analysis. *Am J Health Promot*, 17(1):1–6, ii, 2002.

[303] S. A. French. Pricing effects on food choices. *J Nutr*, 133(3):841S–843S, 2003.

[304] Y. C. Wang, S. L. Gortmaker, A. M. Sobol, and K. M. Kuntz. Estimating the energy gap among us children: a counterfactual approach. *Pediatrics*, 118(6):e1721–33, 2006.

[305] M. Nestle. *Food politics: how the food industry influences nutrition and health.* Berkeley: University of California Press., 2002.

[306] A. Yngve. Food and drink marketing to children: a continuing scandal. *Public Health Nutrition*, 10(10):971–972, 2007.

[307] M. K. Lewis and A. J. Hill. Food advertising on british children's television: a content analysis and experimental study with nine-year olds. *Int J Obes Relat Metab Disord*, 22(3):206–14, 1998.

[308] S. A. Rivkees. Advertised calories per hour...2000+: anti-obesity announcements per hour...0. *J Pediatr Endocrinol Metab*, 20(5):557–8, 2007.

[309] A; Matthews, J; Longfield, and C; Powell. The marketing of unhealthy food to children in europe a report of phase 1 of the children, obesity and associated avoidable chronic diseases project. Technical report, 2005.

[310] G; Hastings, L; McDermott, K; Angus, M; Stead, and S; Thomson. The extent, nature and effects of food promotion to children: A review of the evidence. geneva, switzerland: World health organization, 2006. 2006.

[311] M.G Wootan. *Pestering parents: how food companies market obesity to children*. Center for Science in the Public Interest, 2003.

[312] Martin Caraher, Jane Landon, and Kath Dalmeny. Television advertising and children: lessons from policy development. *Public Health Nutrition*, 9(05):596–605, January 2007.

[313] Ofcom statement on the television advertising of food and drink products to children., 2007.

[314] Corinna Hawkes. Regulating and litigating in the public interest: regulating food marketing to young people worldwide: trends and policy drivers. *American journal of public health*, 97(11):1962–73, November 2007.

Appendix A

Introduction

A.1 Review of Obesity Policy Interventions

A.1.1 Information Interventions

The aim of information interventions is to allow consumers to make an educated choice via provision of the information required to understand the consequences of that choice. Effective information interventions therefore rely on consumers initially unaware of the consequences of specific lifestyle actions changing their habits in response to newly available information.

Food Labelling Legislation

UK law requires that ingredients are listed on pre-packaged foods in weight order. Although many manufacturers list nutrition information, this is not a legal requirement unless making a health claim or adding vitamins or minerals [217]. In 2005, the Food Standards Agency (FSA) introduced a food labelling scheme using traffic lights to indicate the health value of foods.

Participation in the FSA MTL scheme is voluntary- the scheme was compromised in 2006 by the withdrawal of several major companies represented by the Food and Drink Federation (FDF, www.fdf.org.uk) in favour of monochrome percentages of Guideline Daily Amounts (%GDA) [218]. Beard and colleagues argue that if the labelling scheme is not mandatory, effectiveness is severely limited due to the clear conflict of interest of the food industry [219].

The multiple traffic lights (MTL) system was chosen by surveying 2,600 individuals to identify the best understood approach [220]. However, this is not necessarily the approach that will result in the greatest shift in shopping habits. Although some studies have examined understanding, use [221], and recognition [222] of food labels, there is very limited evidence regarding the effectiveness of any labelling scheme to alter food purchasing behaviour. Some supermarkets have re-

leased data on sales patterns of specific products following the introduction of the MTL system [223], but is not of use from a public health perspective. Red warning labels may encourage consumers to buy an alternative product that is not labelled. The FSA MTL scheme is a well established intervention but its effect on total energy intake and energy balance is completely unknown. Research needs to focus on impact on energy balance and how this differs between labelling approaches and different groups.

Mandatory Nutrition Labelling in Restaurants

Although some legislation exists for the provision of nutritional information on packaged food, food purchased for immediate consumption is exempt. Food consumed outside the home is usually less healthful than home prepared food [224] and U.S levels of consumption and portion sizes have risen sharply in recent decades [225, 226]. The subsequent increase in calorific uptake from such meals is a significant contributor to obesity [226]. A frequent policy suggestion is the mandatory provision of nutrition information in restaurant chains, to allow consumers to make more informed choices [227]. A shift in consumer preference towards healthier choices is likely to result in recipe adjustments and menu changes. The benefits of such an approach depend wholly on the degree to which the information provided would influence consumer choice; this remains largely unknown.

Consumers have been consistently shown to substantially underestimate the calorific, fat, trans-fat and sodium content of unhealthful restaurant food [228, 229]. Burton and colleagues [228] carried out a study where individuals were asked to evaluate healthfulness and likelihood of purchase of various common restaurant foods before and after receiving nutrition information. There was a significant shift towards healthier food choices following receipt of information. The authors suggest that provision of nutrition information in restaurants could have an effect on consumption of less healthful foods through both consumer choice and menu changes, and therefore public health. However this study provided information directly to an educated group. Observed effects are unlikely to be applicable to the whole population.

Currently, several restaurant chains voluntarily provide nutritional information [230], but information is not always easily accessible [230–232], or easy to understand [233]. The benefits of leaflet and web information to public health are unclear, but it is unlikely that it plays a significant role in food choice, with concern regarding levels of use and understanding [233], particularly in less educated

groups [221]. An additional problem may be the lack of consumer understanding of what constitutes a high energy diet [234]. Education is likely to be an important component of any successful initiative. Survey data has suggested that food choices may be influenced by available information [235, 236], but there is an absence of data looking directly at food choices in response to nutritional information.

Some empirical evidence does exist for the effect of promotion of healthier alternatives in vending machines. Bergen and Yeh [237] showed that high visibility information labels and motivational posters yielded a significant shift towards healthier soft drink choices in a US college setting. Similar interventions have been shown to be effective in food snack vending machines [238–240]. However individuals may compensate by purchasing high energy snacks elsewhere; impact on overall consumption and energy balance is unknown.

The effect of provision of nutrition in restaurants and on vending machines is likely to depend greatly on the clarity and prominence of information [241] and the receptiveness of the individual consumer [221, 242]. Positioning calorific information at the point of sale with equal prominence to the price is likely to have a much more significant effect than information sheets, which the majority do not consult [241]. Those that do are likely to be those least at risk [221]. Current voluntary provision of information by restaurant chains is unlikely to have a significant effect on customer choice, and any proposed legislation to increase the availability of nutritional information is likely to be met with opposition from the food industry [243]. Implementation of this legislation is likely to be costly, both for government and industry [243] and there is no evidence that consumer eating habits will be altered. Research is needed into real effects of information provision, particularly in high risk groups.

Signs Encouraging Physical Activity

Some studies have measured the effect of point-of-decision prompts to increase stair use when lifts and escalators are available. These take the form of signs that highlight the number of calories burnt or promote the general benefits of exercise, and have been shown to be effective in increasing stair use in several groups [244–247]. Although no evidence exists for effect on overall activity levels or energy balance, the low cost of intervention may make sign introduction viable in settings such as shopping centres and office buildings.

Teaching Healthy Living in the Classroom

Several bodies and individuals, most notably the House of Commons Select Committee in 2004, have suggested that healthy living education should be part of the national curriculum:

> As well as practical cookery lessons and classroom lessons about nutrition, children should also be taught how to understand food labelling and how to distinguish food advertising and marketing from objective fact [15].

Although numerous short term nutrition education interventions have been carried out, mainly to promote weight loss, there are insufficient long term studies to gauge the public health effects of nutrition classes in a school setting [248]. The benefits of equipping children with a good comprehension of what constitutes healthy diet, and the ability to prepare nutritious balanced meals, may only be realised in adulthood [249]. Kahn and colleagues review the evidence available for the effects of education programmes to increase physical activity levels amongst schoolchildren, but are unable to draw any conclusions [250].

Mass Marketing of a Healthy Lifestyle

This approach relies on using mass marketing techniques to deliver a health message to the population, and suggested as early as 1970 [251]. The UK NHS 'Just Eat More: 5-a-day' campaign is one of the most prominent, although lacks firm evidence of an increase in uptake [252]. However similar campaigns on a smaller scale can be effective, at least in the short term [253]. Despite reported success in terms of awareness, it is difficult to estimate health benefits of these schemes.

Marketing campaigns have also been implemented to increase rates of physical activity, mostly in children [254]. The VERB campaign [254], is a commercial marketing strategy, set up the aim with the aim to promote physical activity in American 9-13 year olds, with reported success. Again, figures are not easily translatable into quantifiable population health benefits.

A.1.2 Accessibility Interventions

Accessibility interventions attempt to reduce the obesogenic environment by making poor health choices less convenient; and good health choices easier.

Banning Vending Machines in Schools

UK and US schoolchildren generally have extensive access to vending machines, predominantly selling sugar sweetened beverages (SSBs) and snack foods [255, 256]. These items are also aggressively marketed to children [257]. SSB consumption has been associated with increased risk of obesity [258], and therefore banning placement in schools has been recommended to reduce consumption levels [259]. Vending machines are currently banned in French schools and there has been support for similar legislation in the US and UK.

Use of school vending machines has been associated with increased intake of SSBs [260], and a behavioural intervention to dramatically reduce SSB consumption reported significant improvement in participant BMI [261]. Despite this, there is still no solid evidence that removal of vending machines will result in a decrease in energy intake, as children may compensate with higher intake elsewhere, although the presence of machines indicate that school placement generates extra sales. Dependence of some schools on money generated by vending contracts further complicates matters [243].

Improving access to healthy choices in schools

Children spend a considerable proportion of their time in school, and meals eaten in school compose a substantial component of energy and nutrient intake. Despite compulsory provision of healthy foods in schools [262] children's food choices remain worse than those made at home [263]. Children in UK schools have access to healthy options, but uptake is low. There is some evidence to suggest that children from more deprived backgrounds make less healthy choices [264]. Students with greater nutritional awareness also tend to make better choices [265]. Offering high fat menu items less frequently may result in lower fat intake [263]; effects of removing, limiting or rationing unhealthy choices are unclear. Packed lunches also typically provide few nutrients and are high in sugars and fats [266]. Subsequently, education for parents encouraging provision of a healthier packed lunch has been recommended [266]. Some schools operate a 'healthy packed lunch policy', but no studies have examined impact on calories and nutrients consumed.

Improving access to healthy foods

Individuals from lower socio-economic groups tend to have inferior diets, with higher consumption of fats and sugars, and lower intake of fruits and vegeta-

bles [267]. This has been partially ascribed to the existence of 'food deserts'; areas where residents do not have access to healthy affordable food [268]. UK Government research has recommended improved access to healthy foods to reduce health inequalities [268]. Outside the US however, evidence of the existence of food deserts is limited; larger, more recent UK studies have indicated that accessibility of healthy food is not a limiting factor on diet quality [269, 270]; rather, price and education appear to be more relevant factors [63].

Location of Fast Food Restaurants

Fast food restaurants are much more common in deprived than affluent areas [63, 271, 272]. There are several plausible reasons for this, as Pearce and colleagues point out [271]; it is not clear if or how this distribution contributes to diet disparity between rich and poor. High availability and consumption of fast food is likely to be a contributory factor in the obesity epidemic [226]. City planners have suggested limitations on fast food outlets near schools and other youth orientated areas, and the density of fast food outlets both per capita and by geographical area [273]. However, the effects of such interventions have not been estimated.

Legislation on food quality

Introducing codes to set formal limits on energy density would compel manufacturers to develop healthier recipes. New York has imposed a ban on the use of trans fats in restaurant food [274]. However, this intervention is aimed at reducing levels of heart disease associated with consumption of trans fats [275] rather than lowering energy intake. No plans are in place to impose legislation on food quality to curtail obesity, with governments preferring to allow the food industry to self-regulate [15].

Altering the Built Environment

Lower levels of physical activity are associated with risk of obesity [276]. Currently many environments do not encourage physical activity with few or no amenities within walking distance, areas that are not safe at night [277], lack of cycle paths and sporting facilities [103], lifts more accessible than stairs [278], and a culture of car use [279].

The built environment is well established as a contributory factor to obesity [277, 280]. Several studies have suggested that such considerations should be

borne in mind when building new neighbourhoods or changes made to existing ones [281, 282]. However, evidence for likely efficacy of interventions is difficult to generate [283] and few studies have carried out interventions by changing the built environment. Stokes and colleagues estimated the potential health benefits from increased walking associated with provision of a rail transit system [284].

Although the argument for immediate action is compelling [103] there is no evidence that investment in new sports facilities and cycle paths will have a significant impact on obesity. Studies will need to show that such investment can benefit significant numbers of the population, rather than the active few.

A.1.3 Price Interventions

Economics plays a significant role in many daily decisions, including food purchasing [285]. Artificial adjustment of costs, via taxation or subsidy, may encourage the population to make more optimal health choices.

Taxation of Energy Dense and Unhealthful Food

Substantial evidence has implicated increased consumption of energy dense foods with increased risk of obesity [5]. Taxation of unhealthy and energy dense food has been suggested as method of restricting consumption of energy dense low-nutrient food [249, 286, 287] having proved to be a successful policy in reducing smoking prevalence [288]. Further, revenue raised could be invested into anti-obesity programs and subsidy of healthy foods [289].

Evidence for possible effects of food taxation is extremely limited. Kim and Kawachi [290] describe an association between the presence of a snack/soft drink tax in American states and increase in obesity prevalence between 1991 and 1998, but there is a large likelihood of presence of confounding factors. Kuchler and colleagues [291] investigated price elasticity of savoury snack foods and concluded that demand was relatively unresponsive to price; estimating that a 20% tax on potato chips would result in only a 4-6oz reduction in consumption per capita/year. Effects of taxation systems will be complex and far from uniform across populations and are likely to discriminate against the poor. For realistic policy decisions to be made, research should investigate potential effects in different groups, and across social gradients. Taxation of unhealthy foods, even at low levels can provide substantial income [289], however influence on individual consumption is much less clear. Unintended effects should also be given consideration [290].

Taxation may also raise awareness of unhealthy foods.

Subsidy of Fruit, Vegetables and Other Healthful Food

Use of money raised through taxation of unhealthful food to subsidise healthier food would provide a fiscal incentive to improve eating habits [249, 286]. Fry and Finley [104] have estimated the cost of such an intervention in the European Union, but offer no analysis of effect.

Increased consumption of fruit and vegetables is associated with lower risk of obesity [292] and several other chronic diseases [293]; higher consumption is therefore likely to improve population health. Price and availability may be a limiting factor on fruit and vegetable consumption for some families [294–296]. Although targeted subsidies via food vouchers have been shown to be effective among a small group of low income American women [297, 298], there is an absence of evidence for the potential impact of national subsidy. Additionally, any increase in uptake may not result in health benefits if there is not a commensurate decline in the consumption of unhealthy foods. The influence of a widespread subsidy on consumption and subsequent energy balance is likely to be complex; we are a long way from understanding the viability of this approach.

Subsidy of Healthier Options by Higher Prices of Unhealthful Ones in Controlled Environments

An alternative approach to encouraging healthy eating by price adjustment can be undertaken at schools and workplaces. These populations are almost 'captive' audiences, and it is easy to identify when a healthy choice is made as it directly replaces an alternative [299]. School based initiatives may be particularly effective as healthy eating habits can be established early in life [300, 301], and they have the potential to reach a large proportion of the population [299].

Adjustments in price have been shown to be effective in promoting lower fat options in vending machines [240]. In a school based study raising prices of high fat foods and reducing those of low fat foods resulted in a significant shift in preference towards lower fat choices [302]. Discounting vegetables by 50% increased uptake levels in another study [303]. Assuming no calorie replacement, an alternative menu choice may represent a significant reduction in energy intake. Even a small reduction in daily energy intake may influence weight gain [79, 304]. The potential of such schemes to promote healthier meal choices merit further inves-

tigation, as do the health benefits of those choices at a population level. Studies investigating similar schemes in workplaces and community settings such as hospitals are not available, but are needed to estimate population level effects of price interventions.

Restriction of Deals Appealing to Consumer Sense of 'Value'

The food industry encourages individuals to buy more than they need by appealing to a consumer's sense of perceived value [5]. An additional portion of food is offered at a lower price than the initial portion, examples include 'supersizing' in fast food restaurants and supermarket 'multi-buy' offers. No studies have examined effects of banning such offers.

A.1.4 Marketing

One of the most vociferous campaigns for a policy intervention to reduce obesity is the call to restrict junk food marketing, particularly to children.

Restriction of Advertising Unhealthful Foods to Children

A number of calls have been made to end junk food marketing to children [227, 305] [257, 306]. Food advertisements on children's television are dominated by unhealthful foods of low nutritional value [307, 308]. An estimated £743 million was spent on direct advertising to UK children in 2003 [309].

A study commissioned by the US Centers for Disease Control and Prevention on the influence on food marketing to children and youth drew stern conclusions from a review of 123 studies [257]. A WHO report also found that junk food advertising strongly influenced food choice in children, but could not find evidence of a causal link to obesity [310]. An earlier report by the Center for Science in the Public Interest found that the food marketing industry deliberately attempts to undermine parental food choices, and encourages children to take control of their own diet [311].

There are a broad range of advertising restrictions and voluntary codes in place across developed countries [312]. The UK Office of Communications (Ofcom) has recently introduced legislation banning junk food advertising on programs that have a larger than average proportion of their audience composed of children [313]. Hawkes [314] argues that such legislation is based on ethics, rather than evidence that restriction of advertising will help to prevent obesity. Such evidence is notably

absent, with Hawkes outlining 5 key gaps in current knowledge, most pertinently the lack of evidence of effectiveness of regulation from countries that have had legislation imposed for several years [314]. The food industry has several alternative methods of delivering advertising messages; research needs to find restrictions that can be effective. Nestle provides several other possible approaches, including restrictions or bans on use of cartoon characters and stealth marketing [227]; again however there is no evidence to support the efficacy of these interventions.

A.2 Literature Seach

Searches were carried out in PubMed using MeSH terms to identify papers that used a data-driven approach to identify relationships between multiple variables in epidemiological datasets.

- ("Bayes Theorem"[Majr]) AND "Epidemiology"[Mesh], 22 results. *No relevant results. Mainly models of disease risk.*
- "Artificial Intelligence"[Mesh] AND "Epidemiology" [Mesh], 25 results. *No relevant results. Mainly diagnostic models.*
- "Automatic Data Processing"[Mesh] AND "Epidemiology"[Mesh], 38 results. *No relevant results. Variety of papers around health care systems.*
- "Bayes Theorem"[MAJR] AND "Population Groups"[Mesh], 19 results. *No relevant results. Almost exclusively papers describing prediction of genetic disease risk in population subgroups.*
- "Multivariate Analysis"[MeSH Terms] AND "Epidemiology"[Mesh], 53 results. *No relevant results. Variety of general epidemiological papers.*

Appendix B

Data

B.1 Contents of Health Surveys for England

See figures B.1 and B.2.

B.2 STATA Code for Variable Derivations

B.2.1 Health Surveys for England 2003/6 data

Sex:

```
replace sex=sex-1
lab def sex 0 "c1_Male" 1 "c2_Female",modify
```

Age:

```
gen ageG=88
replace ageG=0 if age>15
replace ageG=1 if age>24
replace ageG=2 if age>34
replace ageG=3 if age>44
replace ageG=4 if age>54
replace ageG=5 if age>64
replace ageG=6 if age>74
label define age_group_lbl  88 "c0to15" 0 "c16to24" 1 "c25to34" \\\
2 "c35to44" 3 "c45to54" 4 "c55to64" 5 "c60to74" 6 "c75plus"
lab val ageG age_group_lbl
```

Dependent Children:

```
gen rel_id=pserial-(hserial*100)
gen depchild=0
local i=1
foreach var of varlist relto01-relto12{
                bys hserial: gen agerel`i'=age if rel_id==`i'
                bys hserial: egen Zagerel`i'=max(agerel`i')
                drop agerel`i'
                replace depchild=1 if `var'>=8 & `var'<=12 & Zagerel`i'<18
                display `i'
                local i=`i'+1
}
drop Zagerel*
label define depchild_lbl 0 "c0_NoDependents" 1 "c1_Dependents"
lab val depchild depchild_lbl
label var depchild "Dependent Child under 18 years"
```

Marital Status:

```
gen currmarried=.
replace currmarried=1 if marital==2
replace currmarried=0 if marital>0 & marital!=2
gen marital_s=.
replace marital_s=0 if currmarried==0 & (couple==2|couple==-1)
replace marital_s=1 if currmarried==1
replace marital_s=1 if currmarried==0 & (couple==1|couple==3)
drop currmarried
```

Health Status:

```
lab def health_lbl 0 "c1_Good" 1 "c2_Fair" 2 "c3_Poor",modify
lab var health_s health_lbl
```

NSSEC (Social Class):

```
gen nssec3=hpnssec3
replace nssec=5 if hpnssec8==99
replace nssec=4 if hpnssec8==8
replace nssec=5 if hpnssec8<0

*use individual social class data if hrp not available
gen temp = 0
replace temp=1 if nssec==5
replace nssec=nssec3 if temp==1
replace nssec=5 if nssec8==99 & temp==1
replace nssec=4 if nssec8==8 & temp==1
replace nssec=5 if nssec8<0 & temp==1
drop temp
replace nssec=cen-1
label define nssec_lbl 0 "c1_ManProf" 1 "c2_Intmdt" 2\\\
"c3_RtnMan" 3 "c4_LTunplyd" 4 "c5_OtherNC",modify
label val nssec nssec_lbl
```

Economic Activity:

```
gen economicAct=econa
recode economicAct(4=2)
replace economicAct=4 if activb==10
replace economicAct=1 if topqual2==8|activb==1
replace economicAct=economicAct-1
lab def economic_lbl 0 "EmployedOrStudent" 1 "UnemployedOrInactive" 2 "Retired" 3 "HomeOrFamily"
lab val economicAct economic_lbl
```

Ethnicity:

```
gen ethnicNW=1
replace ethnicNW=0 if ethinda==1
lab def ethnicNW_lbl 0 "White" 1 "Non-White"
lab val ethnicNW ethnicNW_lbl
```

Education Level:

```
gen educL=.
replace educL=0 if topqual2==1|topqual2==2
replace educL=1 if topqual2>=3
replace educL=2 if topqual2==7
replace educL=3 if topqual2==8|activb==1
lab def educL_lbl 0 "UK Higher Ed." \\\
1 "Below Higher Ed" 2 "None" 3 "Current Student"
lab val educL educL_lbl
```

Leisure/Transport Access:

```
gen transport=.
gen leisure=.
lab def transprt   0 "agree/disagree" 1 "strongly disagree",modify
foreach var of var transprt leisure{
        gen temp=1 if `var'==4
        replace `var'=0
        replace `var'=1 if temp==1
        lab val `var' transprt
        drop temp
}
```

Recreational Physical Activity:

```
gen recpa_g4=hrssptg
recode recpa_g4 (4 5 = 3)
lab def recpa_g4 0 "c0None" 1 "c0to1" 2 "c1to3" 3 "c3plus",modify
```

Incidental Physical Activity:

```
gen inc_pa=daywlk
replace inc_pa=0 if wlk15==2|wlk5int==2
replace inc_pa=1 if daywlk>2
replace inc_pa=2 if daywlk>10
replace inc_pa=3 if daywlk>20
lab def inc_pa_lbl 0 "2 or less" 1 "10 or less" 2 "20 or less" 3 "Above 20"
lab val inc_pa inc_pa_lbl
```

Occupational Physical Activity:

```
gen occ_pa4=workact
replace occ_pa4=3 if workact==4
replace occ_pa4=occ_pa+1 if (hwrklist==1|gardlist==1) & (hevyh!=1&manwork!=1)
replace occ_pa4=occ_pa+2 if hevyh==1|manwork==1
recode occ_pa4(-6=-8)
replace occ_pa4=occ_pa-1 if occ_pa>0
lab def occ_pa_lbl -8 "Refused" -1 "item N/A" 0 "Inactive" 1 "Low Activity" 2 "Moderate" 3 "Active"
lab val occ_pa4 occ_pa_lbl
recode occ_pa4 (4=3)
```

Fried food/Snack/Cake intake:

```
lab def intake_lbl 0 "6 or more times a week" 1 "3-5 times a week" 2 \\\
        "1-2 times a week" 3 "less than once a week" 4 "Rarely/Never",modify
foreach var of var friedfdb snack cakesc{
        gen `var'_itk_4 =`var'-1 if `var'>0
        gen `var'_itk_5 =`var'-1 if `var'>0
        *Reduce levels by 1
        recode `var'_itk_4 (4=3)
        lab val `var'_itk_4 intake_lbl
        lab val `var'_itk_5 intake_lbl
}
*Fried is also put into a 3 group variable
gen friedfdb_itk_3=friedfdb_itk_4-1
recode friedfdb_itk_3(-1=0)
lab def fried_red 0 "3 or more times a week" 1 "1-2 times a week" 2 "less than once a week",modify
lab val friedfdb_itk_3 fried_red
```

Fruit/vegetable intake:

```
gen ftvg_itk_4=porftvg
recode ftvg_itk_4 (1=0) (2 3=1) (4 5=2) (6 7 8 9 =3)
lab def ftvg_lbl 0 "Less than 1" 1 "1 to 3" 2 "3 to 5" 3 "5 or more"
lab val ftvg_itk_4 porftvgg_lbl
```

Body Mass Index:

```
gen bmi_6 = bmivg6-1
lab def bmivg6 0 "below 19" 1 "19 to 24.9" 2 "25 to 29.9"\\\
        3 "30 to 34.9" 4 "35 to 39.9" 5 "40 +",modify
lab val bmi_6 bmivg6
```

Waist-Hip Ratio:

```
gen male_whr_g = menwhgp-1
gen female_whr_g = womwhgp-1
lab def menwhgp 0 "less than 0.80" 1 "0.80, less than 0.85" 2 "0.85, less than 0.90" 3 "0.90,\\\
 less than 0.95" 4 "0.95, less than 1.00" 5 "1.00 or more",modify
lab def womwhgp 0 "less than 0.70" 1 "0.70, less than 0.75" 2 "0.75, less than 0.80" 3 "0.80,\\\
 less than 0.85" 4 "0.85, less than 0.90" 5 "0.90 or more",modify
```

Alcohol intake:

```
gen alcohol_itk_3=dnoft2
recode alcohol_itk_3 (1 2 = 1) (3 4 = 2) (5 6 7 8 = 3)
replace alcohol_i=alcohol_i -1 if alcohol_i>0
lab def alcohol_lbl 0 "Heavy" 1 "Moderate" 2 "Light"
lab var alcohol_itk_3 alcohol_lbl
```

Smoking status:

```
*Smoking
gen smoking_s = cigsta3
replace smoking_s = smoking-1 if smoking_s>0
lab def smoking_lbl 0 "current cigarette smoker" 1 "ex-regular cigarette smoker"\\\
 2 "never regular cigarette smoker",modify
lab val smoking_s smoking_lbl
```

Period status:

```
gen period_s=period-1
lab def period_lbl 0 "Yes" 1 "No",modify
lab val period_s period_lbl
```

Age*:

```
gen cen_ageG=88
replace cen_ageG=0 if age>=16
replace cen_ageG=1 if age>=20
replace cen_ageG=2 if age>=30
replace cen_ageG=3 if age>=40
replace cen_ageG=4 if age>=50
replace cen_ageG=5 if age>=65
label define cen_age_group_lbl  0 "c16to19" 1 "c20to29" 2 "c30to39" 3 "c40to49" 4 "c50to64" \\\
 5 "c65to74" 6 "c75plus"
lab val cen_ageG cen_age_group_lbl
```

Social Status:

```
gen cen_socgrade=schrpg4
recode cen_socgrade (3=2) (4=3)
replace cen_socgrade=cen_socgrade-1
replace cen_socgrade=3 if hpnssec8==8
replace cen_socgrade=4 if cen_soc<0
lab def socgrade_lbl 0 "Managerial" 1 "Manual" 2 "Semi/unskilled" 3 "unemployed" 4 "unknown"
lab val cen_soc socgrade_lbl
```

Economic Activity*

```
recode cen_econa4 (3=4)
replace cen_econa4=3 if topqual2==8|activb==1
replace cen_econa4=cen_econa-1
lab def cen_econact4_lbl 0 "Employed" 1 "Unemployed" 2 "FT Student" 3 "Other Economically inactive"
lab val cen_econa4 cen_econact4_lbl
```

B.2.2 Census 2001 data

cSex:

```
gen csex=sex
recode csex(2=0)
lab def sex_lbl 0 "Male" 1 "Female"
lab val csex sex_lbl
```

cAge:

```
gen cen_ageG=age
recode cen_age (25=20) (60=50)
*Recode to numeric categories
recode cen_age (16=0) (20=1) (30=2) (40=3) (50=4) (65=5)
lab def cen_age_lbl 0 "16-19" 1 "20-29" 2 "30-39" 3 "40-49" 4 "50-64" 5 "65-74",modify
lab val cen_age cen_age_lbl
```

cDependent Children:

```
gen depchild = 0
replace depchild=1 if (fndepcha==1) & (relto==1|relto==2|relto==3) & (gen!=1)
```

cMarital Status:

```
gen marital_s = 0
replace marital_s = 1 if (famtyp==2|famtyp==3) & (relto==1|relto==2|relto==3) & (gen==2)
lab def marital_s_lbl 0 "Single" 1 "Couple",modify
lab val marital_s marital_s_lbl
```

cHealth status:

```
gen chealth=health-1
lab def health_lbl 0 "Good" 1 "Fairly Good" 2 "Poor"
lab val chealth health_lbl
```

cSocial Status:

```
gen socgrade=hrsoc
recode socgrade (1=0) (2 3=1) (4=2) (5=3) (-9=4)
lab def socgrade_lbl 0 "Managerial" 1 "Manual" 2 "Semi/unskilled" 3 "unemployed" 4 "unknown"
lab val socgrade socgrade_lbl
```

cEconomic Activity:

```
gen econact=econach
replace econact=econact-1
replace econact = 2 if student==1
lab def econa_lbl 0 "Employed" 1 "Unemployed" 2 "FT Student" 3 "Other Economically inactive"
lab val econact econa_lbl
tab econact
```

cEthnicity:

```
gen ethnicity=.
replace ethn=0 if ethewa>0
replace ethn=1 if ethewa>3
lab def ethnicity_lbl 0 "White" 1 "Non-White"
lab val ethnicity ethnicity_lbl
```

cEducation Level:

```
gen EducationL=qualv
recode EducationL (1=2) (6 4 3 2 =1) (5=0)
replace EducationL=3 if student==1
lab def EducL_lbl 0 "UK Higher" 1 "Some" 2  "None" 3 "Current Student"
lab val EducationL EducL_lbl
```

Health Survey for England 2003: Contents

Household data

Household size, composition and relationships
Accommodation tenure and number of bedrooms
Economic status/occupation of Household Reference Person
Household income
Type of dwelling and area
Car ownership
Smoking in household

Individual level information

	Age										
	0-1	2-3	4	5-6	7	8-10	11-12	13-15	16-34	35-64	65+
Interviewer visit											
General health, longstanding illness, limiting longstanding illness, acute sickness, fractures	●	●	●	●	●	●	●	●	●	●	●
CVD, including use of services									●	●	●
Rose Angina questionnaire									●	●	●
Physical activity									●	●	●
Smoking						●[a]	●[a]	●[a]	●[b]	●	●
Drinking (seven day period)						●[a]	●[a]	●[a]	●[b]	●	●
Fruit and vegetable consumption				●	●	●	●	●	●	●	●
Economic status/occupation, educational achievement									●	●	●
Ethnic origin	●	●	●	●	●	●	●	●	●	●	●
Parental health									●	●	●
Height measurement		●	●	●	●	●	●	●	●	●	●
Weight measurement	●	●	●	●	●	●	●	●	●	●	●
Cycling safety						●[a]	●[a]				
Psychosocial health (GHQ 12)								●[a]	●[a]	●[a]	●[a]
Euroqol general health (EQ5D)									●[a]	●[a]	●[a]
Social support, social capital									●[a]	●[a]	●[a]
Use of contraceptive pill									●[a]	●[a]	●[a]
Hormone replacement therapy									●[c]	●[a]	●[a]
Nurse visit											
Prescribed medicines and vitamin supplements	●	●	●	●	●	●	●	●	●	●	●
Nicotine replacements									●	●	●
Immunisations	●										
Blood pressure				●	●	●	●	●	●	●	●
Waist and hip circumference									●	●	●
Blood sample – total & HDL cholesterol, fibrinogen, c-reactive protein, glycated haemoglobin									●	●	●
Saliva sample – cotinine			●	●	●	●	●	●			
Infant length	●										
Eating habits (fat, salt)									●[a]	●[a]	●[a]
Additional nurse procedures in the sub-sample (extended nurse visit)											
Saliva (cotinine)									●	●	●
Fasting blood samples – triglycerides, LDL cholesterol, glucose										●	●
Urine sample									●	●	●

[a] These modules were administered by self completion.

[b] This module was administered by self-completion for those aged 16-17 and some aged 18-24.

[c] 18+ only (there are no HRT questions in the young adult self-completion).

Figure B.1: Summary of Health Survey for England 2003 contents

Figure A

Health Survey for England 2006: Contents

Household data
Household size, composition and relationships
Accommodation tenure and number of bedrooms
Economic status/occupation of Household Reference Person
Household income
Smoking in household
Type of dwelling and area
Car ownership

Individual level information	**Age**								
	0-1	2-3	4	5-7	8-10	11-12	13-15	16-64	65+
Interviewer visit									
General health, longstanding illness, limiting longstanding illness, acute sickness, fractures	●	●	●	●	●	●	●	●	●
Use of social care services									●
Carers' responsibilities								●	●
CVD								●	●[a]
Child Physical activity		●	●	●	●	●	●		
Adult Physical activity (short version)									●[a]
Adult Physical activity (long version)								●	●[a]
Smoking					●[b]	●[b]	●[b]	●[c]	●
Drinking (seven day period)					●[b]	●[b]	●[b]	●[c]	●
Fruit and vegetable consumption (and salt)				●	●	●	●	●	●
Eating habits (fat, sugar)		●	●	●	●	●	●		
Economic status/occupation, educational achievement								●	●
Ethnic origin	●	●	●	●	●	●	●	●	●
Height measurement		●	●	●	●	●	●	●	●
Weight measurement	●	●	●	●	●	●	●	●	●
Reported birth weight	●	●	●	●	●	●	●		
Cycling safety					●[b]	●[b]			
Psychosocial health (GHQ 12)							●[b]	●[b]	●[b]
Euroqol general health (EQ5D)								●[b]	●[b]
Social support, social capital								●[b]	●[b]
Strengths and difficulties			●[d]	●[d]	●[d]	●[d]	●[d]		
Perception of weight					●[b]	●[b]	●[b]		
Use of contraceptive pill								●[b,e]	●[b,e]
Hormone replacement therapy								●[b,e]	●[b,e]
Nurse visit									
Prescribed medicines and vitamin supplements	●	●	●	●	●	●	●	●	●
Nicotine replacements								●	●
Immunisations	●								
Blood pressure				●	●	●	●	●	●
Eating habits								●[b]	●[b]
Infant length	●								
Waist and hip circumference						●	●	●	●
Demi-span									●
Blood sample – total & HDL cholesterol, ferritin, haemoglobin, glycated haemoglobin, fibrinogen, c-reactive protein								●	●
Saliva sample (cotinine)			●	●	●	●	●		
Urine sample								●	●

[a] To avoid an overlong interview for informants aged 65 and over, they were randomly allocated to one of two groups: one group answered CVD questions and a short version of the physical activity questions, and the second group completed the full version of the physical activity questions, but not the CVD questions.
[b] These modules were administered by self-completion.
[c] This module was administered by self-completion for those aged 16-17 and some aged 18-24.
[d] This module was administered by self-completion to parents of 4-15 year olds.
[e] This is asked of women aged 18 and over only (there are no HRT questions in the young adult self-completion).

Figure B.2: Summary of Health Survey for England 2003 contents

Appendix C

Software Development

C.1 Provision of C# Code of the Implementation of Metropolis Hastings Sampling over the Space of Bayesian Network Topologies

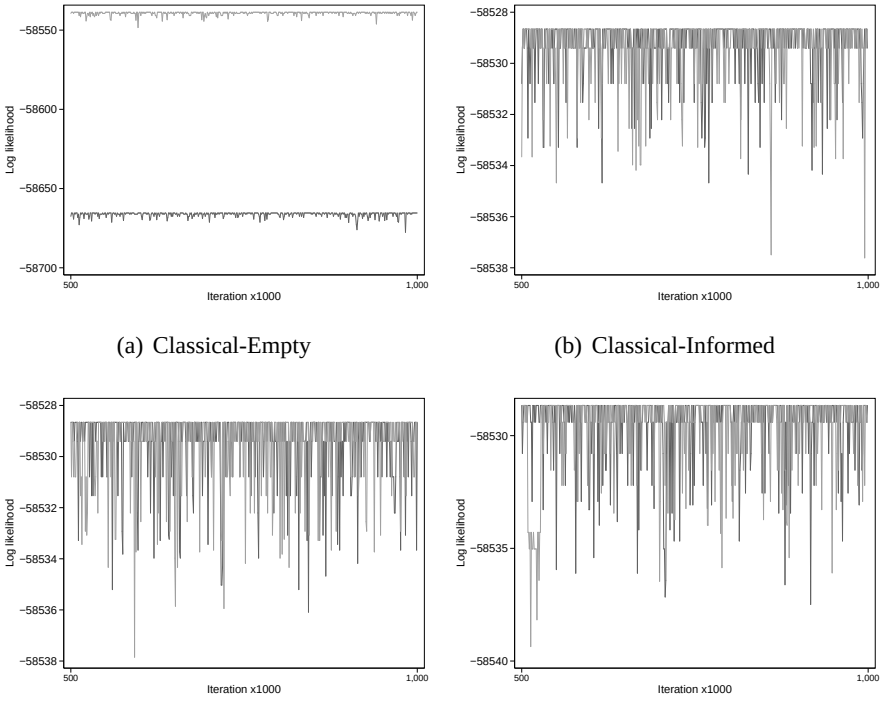
Code is available as a Microsoft Visual Studio solution in the form of a *.rar* archive from `http://personalpages.manchester.ac.uk/postgrad/nicholas.harding/njharding_thesis_samplingoverBNtopologies.rar`, or alternatively via `http://tinyurl.com/nhardingthesis2011`.

C.2 Evidence traces from evaluation of Grzegorczyk-Husmeier move

See figure C.1.

C.3 Evidence traces from evaluation of Multiple Reversal move

See figure C.2.



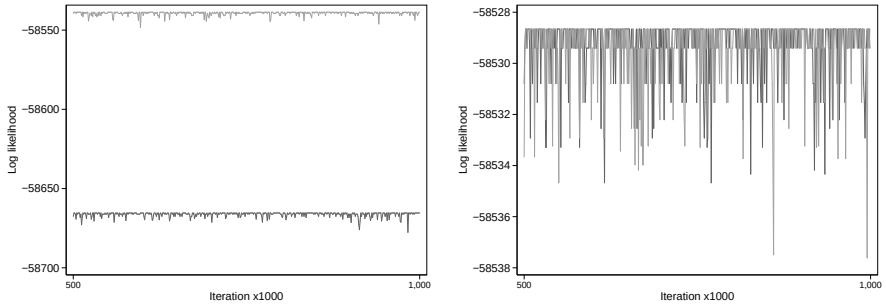
(a) Classical-Empty

(b) Classical-Informed

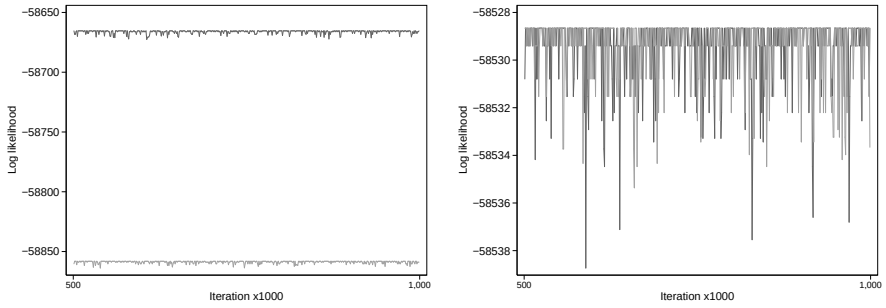
(c) REV-Empty

(d) REV-Informed

Figure C.1: Evidence traces to compare convergence of Markov chain between schemes: REV vs Classical



(a) Classical-Empty (b) Classical-Informed



(c) MR-Empty (d) MR-Informed

Figure C.2: Evidence traces to compare convergence of Markov chain between schemes: REV vs Classical

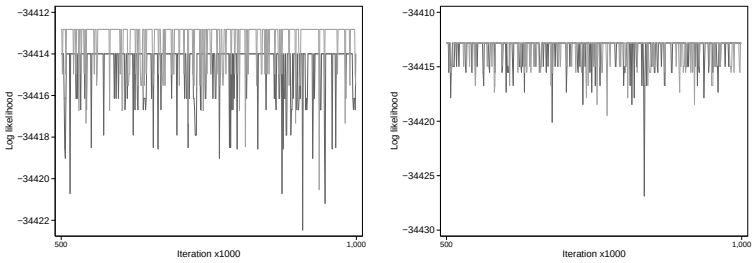
Appendix D

Combination of High and Low Resolution Datasets Using Bayesian Networks

D.1 Mixing of Markov Chain During Metropolis Hastings Sampling

D.1.1 Males

See figures D.1 and D.2.

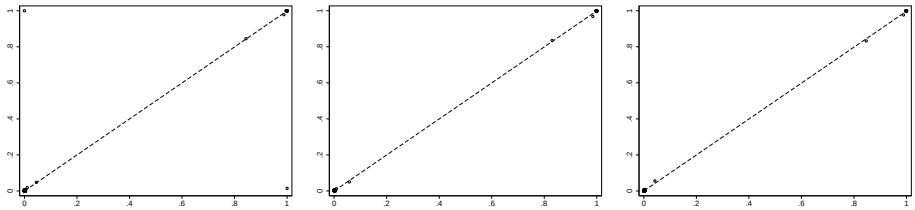


(a) empty/empty (b) informed/informed

Figure D.1: Evidence traces of Markov chain during Metropolis Hastings sampling of Bayesian network topologies modelling influence of socio-demographic variables on health behaviours (males)

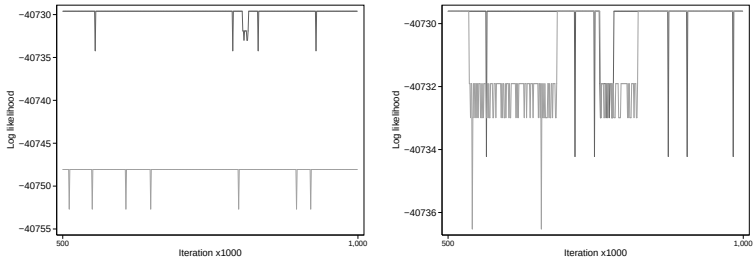
D.1.2 Females

See figures D.3 and D.4.



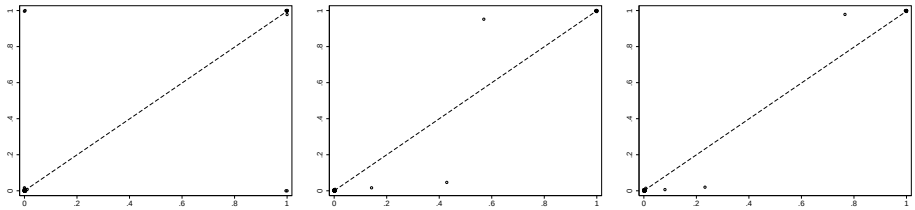
(a) empty-empty (b) informed-informed (c) empty*-informed

Figure D.2: Scatter plots of edge relation features following Metropolis Hastings sampling of Bayesian network topologies modelling influence of socio-demographic variables on health behaviours (males)



(a) empty/empty (b) informed/informed

Figure D.3: Evidence traces of Markov chain during Metropolis Hastings sampling of Bayesian network topologies modelling influence of socio-demographic variables on health behaviours (females)



(a) empty-empty (b) informed-informed (c) empty*-informed

Figure D.4: Scatter plots of edge relation features following Metropolis Hastings sampling of Bayesian network topologies modelling influence of socio-demographic variables on health behaviours (females)

D.2 R Script: Functions

```
#EE tool functions
library("Rgraphviz")
library("fSeries")

getDataFile<-function(fileName,filePath){
        fullPath<-paste(filePath, fileName, sep="")
        output<-read.table(fullPath,header = TRUE)
}

getRelParents<-function(NOI,g1){
        y<-inEdges(NOI,g1)
        temp<-temp<-paste(rep(" ",0), collapse="")
        for(i in 1:length(y)){
                temp<-append(y[i][[1]],temp)
        }
        relParents<-unique(temp[temp!=""])
}

fillGroupInputs<-function(relParents){
        nGroups<-prod(nlevels[relParents])
        gInputs<-matrix(0,nGroups,length(relParents),FALSE,list(1:nGroups,relParents))

        for(parent in 1:length(relParents)){
                ID<-relParents[parent]
                levels<-nlevels[ID][[1]]

                #product of levels of rightmost
                rWeight<-1
                x<-(parent+1)
                while(x<=length(relParents)){
                        rWeight<-rWeight*nlevels[relParents[x]][[1]]
                        x<-x+1
                }
                rWeight

                #product of levels of leftmost
                lWeight<-1
                x<-(parent-1)
                while(x>0){
                        lWeight<-lWeight*nlevels[relParents[x]][[1]]
                        x<-x-1
                }
                lWeight
                index<-1
                #outerloop represents product of left nodes
                for(i in 1:lWeight){
                        counter<-0
                        for(j in 1:levels){
                                #inner loop represents repeats
                                for(k in 1:rWeight){
                                        gInputs[index,ID]<-counter
                                        index<-index+1
                                }
                                counter<-counter+1
                        }
                }
        }
        output<-gInputs
}

genPopCounts<-function(inputs,censusData){
#inputs refers to the definitions of the groups, returns counts in each group.
output<-0
        for(i in 1:nrow(inputs)){
                #takes account of thefrquency weighting!! Hence much faster!
                output[i]<-sum((censusData[apply(censusData[names(inputs[i,])],1,function(x)
                                {all(x==inputs[i,])}),])\$X_freq)
        }
        pop<-output
```

```
}

genPopCountsB<-function(inputs,censusData){
output<-0
        for(i in 1:nrow(inputs)){
                #no account of the frquency weighting!!
                output[i]<-nrow((censusData[apply(censusData[names(inputs[i,])],1,function(x)
                                {all(x==inputs[i,])}),]))
        }
        pop<-output
}

genhseCounts<-function(inputs,data,nointerest,nlevels){
#inputs- input levels
#data- dataset to query
#nointerest- the node id of interest
#nlevels- number of levels of this node
output<-NULL
        cNames<-colnames(inputs)
        for(i in 1:nrow(inputs)){ #loop each row of inputs
                temp<-data[apply(data[cNames],1,function(x) {all(x==inputs[i,])}),][nointerest]
                output<-rbind(output,tabulate(as.matrix(temp)+1,nlevels))
        }
        pop<-output
}

graphFromArcList<-function(arcList,nNodes,nodeLabels){

        #declare matrix
        adjmatrix<-matrix(0,nrow=nNodes,ncol=nNodes)
        x<-strsplit(arcList," ", extended = TRUE, fixed = FALSE, perl = FALSE)
        z<-strsplit(x[[1]], "\"", extended = TRUE, fixed = FALSE, perl = FALSE)
        for(i in z){
                a<-as.numeric(i[1])+1
                b<-as.numeric(i[2])+1
                adjmatrix[a,b]<-1
        }

        g1<-new("graphAM", adjmatrix,"directed")
        g1<-as(g1,"graphNEL")  #convert to graphNEL

        if(!missing(nodeLabels)){
                nodes(g1)<-nodeLabels
        }

        nAttrs <- list()
        eAttrs <- list()

        out<-g1
}
##~~PROGRAM~~##
##Input:
## k- number of outcome levels of node of interest
## state- state of interest (i.e. 1-k)
## popCounts- vector of counts of 'real' population in each group
## hseCounts- matrix n x k, counts in each group of each outcome
## truePopSize- size of the population to simulate

library(MCMCpack)
simulate <- function(k,state,popCounts,hseCounts,truePopSize) {

        #define nGroups
        nGroups<-length(popCounts)

        #get Dir estimate of prob in each popn group
        popDir<-rdirichlet(1,popCounts+1)

        #use this to generate estimate of 'true' population
        estPopCounts<-rmultinom(1,truePopSize,popDir)
```

```
        #get HSE Dirichlet from hseCounts, this will be a 'n' x 'k' matrix
        hseDir<-t(apply(hseCounts+1,1,rdirichlet,n=1))

        #using this Dir estimate, generate populations
        ans<-t(mapply(function(x) rmultinom(1,estPopCounts[x],hseDir[x,]),x=1:nGroups))

        #finally count total inds that fall into group of interest
        final<-sum(ans[,state])
}
```

D.3 R Script: Worked example 1

```
rm(list = ls())

#Read in Functions
source("C:\\Documents and Settings\\SPR-User01\\My Documents\\R\\eeToolFunctions.r")
source("C:\\Documents and Settings\\SPR-User01\\My Documents\\R\\simulateFromDir.r")

#Read in HSE
filePath<-"C:\\Documents and Settings\\SPR-User01\\My Documents\\PhD Obesity Epidemiology
        \\HSE and Census2001 data\\HSE_2006_STATA\\microsim data\\"
hse2006MaleData<-getDataFile("male_microsim_hrssport2006.txt",filePath)
hse2006FemaleData<-getDataFile("female_microsim_hrssport2006.txt",filePath)

#Read in Census
cen2001MaleData<-getDataFile("malesCen2001_G.txt",filePath)
cen2001FemaleData<-getDataFile("femalesCen2001_G.txt",filePath)

#INPUTS:
#census data (from stata)
#DAG representation
#nLevels array
#dirichlet distributions (from HSE)
#vector of nodes of interest, vector of states of nodes.

nodeLabels<-names(hse2006MaleData)
arcListM<-"2~12 2~13 2~14 3~9 4~11 5~8 5~10 5~12 6~13 6~14 7~14 9~8 10~9 10~11"
arcListF<-"2~12 2~13 2~14 3~10 4~12 5~8 6~14 7~11 9~8 10~9 11~10 12~11 14~13"
nNodes<-length(hse2006MaleData)
nlevels<-apply(hse2006MaleData,2,max)+1
malGr<-graphFromArcList(arcListM,nNodes,nodeLabels)
femGr<-graphFromArcList(arcListF,nNodes,nodeLabels)

#APPLICATION 1: RPA=0
NOI<-c(nodeLabels[13])
k<-nlevels[NOI]
state<-1

#Male: Inputs of simulate:
relParentsM<-getRelParents(NOI,malGr)
nGroupsM<-prod(nlevels[relParentsM])
gInputsM<-fillGroupInputs(relParentsM)
popCountsM<-genPopCounts(gInputsM,cen2001MaleData)
hseCountsM<-genhseCounts(gInputsM,hse2006MaleData,NOI,k)
dataM<-data.frame(cbind(gInputsM,popCountsM,hseCountsM))
truePopSizeM<-877100

#Female: Inputs of simulate:
relParentsF<-getRelParents(NOI,femGr)
nGroupsF<-prod(nlevels[relParentsF])
gInputsF<-fillGroupInputs(relParentsF)
popCountsF<-genPopCounts(gInputsF,cen2001FemaleData)
hseCountsF<-genhseCounts(gInputsF,hse2006FemaleData,NOI,k)
dataF<-data.frame(cbind(gInputsF,popCountsF,hseCountsF))
```

```
truePopSizeF <-904754

a<-rep(0,100000)
b<-rep(0,100000)
for(i in 1:100000){
        a[i]<-simulate(k,state,popCountsM,hseCountsM,truePopSizeM)
        b[i]<-simulate(k,state,popCountsF,hseCountsF,truePopSizeF)
}
mean(a+b)
mean(a)
mean(b)

quantile(a,c(0.025,0.975))
quantile(b,c(0.025,0.975))
quantile(a+b,c(0.025,0.975))

mean(a)/truePopSizeM
mean(b)/truePopSizeF
mean(a+b)/(truePopSizeM+truePopSizeF)
```

D.4 R Script: Worked example 2

```
#use non grouped census data
rm(list = ls())

#Read in Functions
source("C:\\Documents and Settings\\SPR-User01\\My Documents\\R\\eeToolFunctions.r")
source("C:\\Documents and Settings\\SPR-User01\\My Documents\\R\\simulateFromDir.r")

#Read in HSE
filePath <-"C:\\Documents and Settings\\SPR-User01\\My Documents\\PhD Obesity Epidemiology
        \\HSE and Census2001 data\\HSE_2006_STATA\\microsim data\\"
hse2006MaleData<-getDataFile("male_microsim_hrssport2006.txt",filePath)
hse2006FemaleData<-getDataFile("female_microsim_hrssport2006.txt",filePath)

#Read in Census
cen2001MaleData<-getDataFile("malesCen2001.txt",filePath)
cen2001FemaleData<-getDataFile("femalesCen2001.txt",filePath)

nodeLabels<-names(hse2006MaleData)
arcListM <-"2~12 2~13 2~14 3~9 4~11 5~8 5~10 5~12 6~13 6~14 7~14 9~8 10~9 10~11"
arcListF <-"2~12 2~13 2~14 3~10 4~12 5~8 6~14 7~11 9~8 10~9 11~10 12~11 14~13"
nNodes<-length(hse2006MaleData)
nlevels<-apply(hse2006MaleData,2,max)+1
malGr<-graphFromArcList(arcListM,nNodes,nodeLabels)
femGr<-graphFromArcList(arcListF,nNodes,nodeLabels)

#Start with occ_pa
NOI<-c(nodeLabels[15])
k<-nlevels[NOI]

parents<-getRelParents(NOI,femGr)
gInputs<-fillGroupInputs(parents)

hseCounts<-genhseCounts(gInputs,hse2006FemaleData,NOI,k)
hseDir<-t(apply(hseCounts+1,1,rdirichlet,n=1)) #samples from dirichlet

#OPTION B- GET COUNTS, THEN DISTRIBUTE!!
censusData<-cen2001FemaleData  #temp holder
censusData<-cbind(censusData,temp=99)
names(censusData)[names(censusData)=='temp'] <- NOI
inputs<-gInputs

popCounts<-genPopCountsB(gInputs,cen2001FemaleData) #sum of census data in each group
```

```
nGroups<-nrow(gInputs)
#multinomial sample for each group
ans<-t(mapply(function(x) rmultinom(1,popCounts[x],hseDir[x,]),x=1:nGroups))
#use ans to generate a vector
for(i in 1:nGroups){
        v<-c(rep(1:ncol(ans),ans[i,])-1) #generates vector based on ans
        #randomly sort vector (using sample)
        y<-sample(v)
        #append vector to censusData as extra column; should be same length!
        censusData[apply(censusData[names(inputs[i,])],1,function(x) {all(x==inputs[i,])}),][NOI]<-y
}

#now we move to the next in the list!
#In this case inc_pa
```

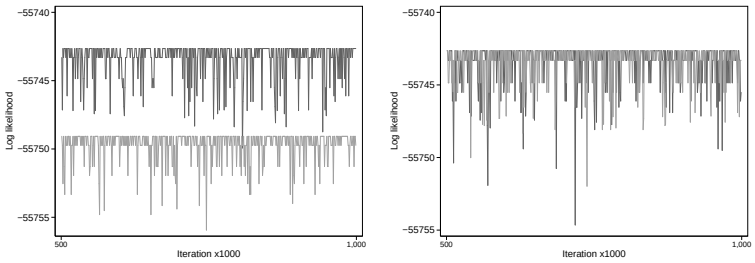
Appendix E

Identification of Predictors of Waist to Hip Ratio in UK Adults

E.1 Mixing of Markov Chain During Metropolis Hastings Sampling

E.1.1 Males

See figures E.1 and E.2.

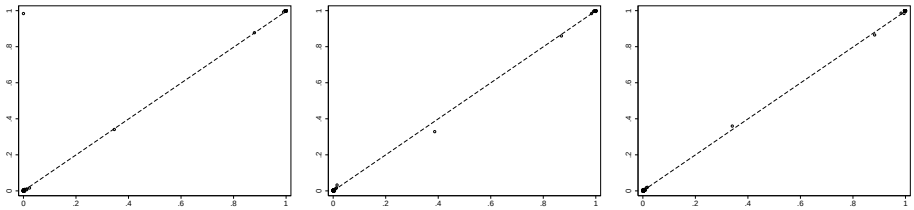


(a) empty/empty (b) informed/informed

Figure E.1: Evidence traces of Markov chain during Metropolis Hastings sampling of Bayesian network topologies modelling factors associated with fat distribution (males)

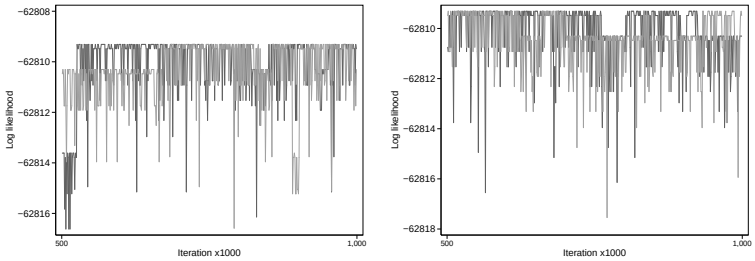
E.1.2 Females

See figures E.3 and E.4.



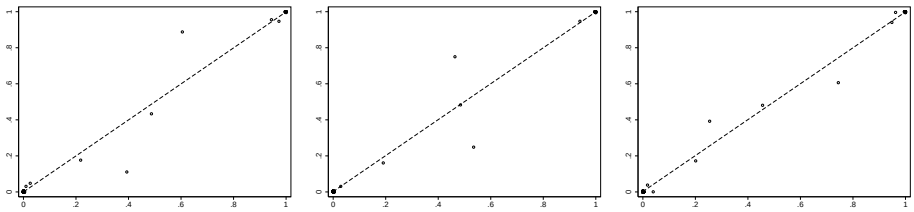
(a) empty-empty (b) informed-informed (c) empty-informed

Figure E.2: Scatter plots of edge relation features following Metropolis Hastings sampling of Bayesian network topologies modelling factors associated with fat distribution (males)



(a) empty/empty (b) informed/informed

Figure E.3: Evidence traces of Markov chain during Metropolis Hastings sampling of Bayesian network topologies modelling factors associated with fat distribution (females)



(a) empty-empty (b) informed-informed (c) empty-informed

Figure E.4: Scatter plots of edge relation features following Metropolis Hastings sampling of Bayesian network topologies modelling factors associated with fat distribution (females)